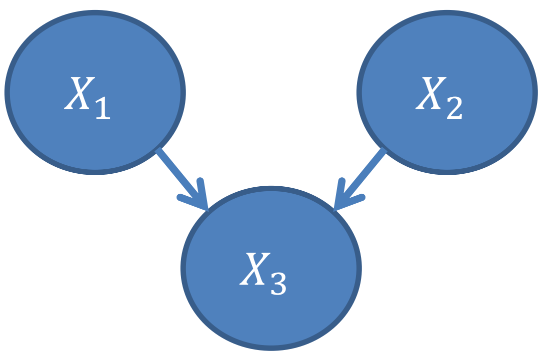
Graphical Event Models and Causal Event Models

Chris Meek
Microsoft Research

Graphical Models

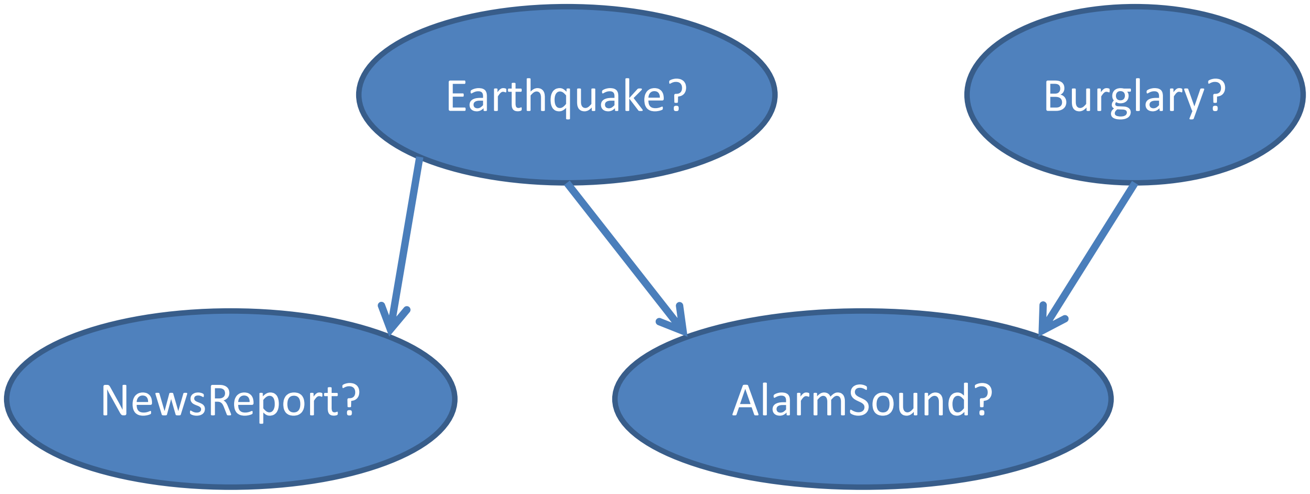
- Defines a joint distribution $P(X)$ over a set of variables $X = \{X_1, \dots, X_n\}$
- A graphical model $\mathcal{M} = < G, \Theta >$
  - $G = < X, E >$ is a directed acyclic graph.
  - $\Theta = \{\Theta_1, \dots, \Theta_n\}$ where Θ_i defines the conditional distribution $P(X_i|\pi_i)$ where π_i are the parents of X_i in G.
- Learning: Assume we see many draws from $P(X)$.



$$P(X_1) = f_1(X_1, \Theta_1)$$
$$P(X_2) = f_2(X_2, \Theta_2)$$
$$P(X_3|X_1, X_2) = f_3(X_3, X_1, X_2, \Theta_3)$$

Graphical Models

- Explaining away type reasoning
  - What is probability of Burglary given AlarmSound?
  - What is probability of Burglary given AlarmSound and a NewsReport of an earthquake?



Graphical Models

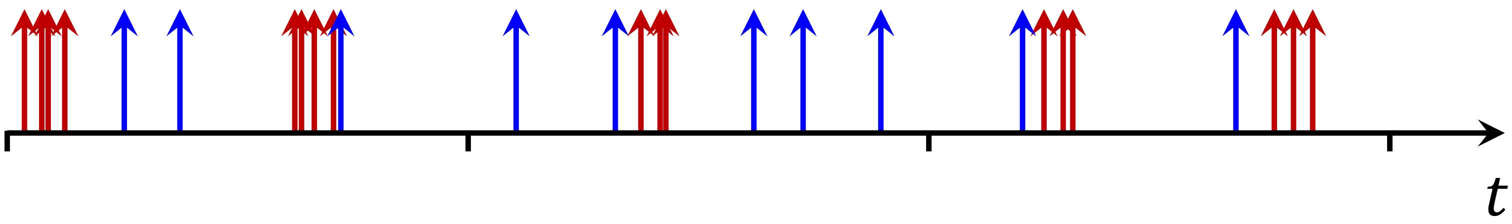
- Explaining away type reasoning
  - What is probability of Burglary given AlarmSound?
  - What is probability of Burglary given AlarmSound and a NewsReport of an earthquake?
  - What if the NewsReport said the earthquake was **after** the Alarm went off?

Outline

- Temporal Event Sequences
- Graphical Event Models
- Learning Graphical Event Models
- Learning Causal Dependencies
  – Causal Event Model ⇔ Graphical Event Models

Modeling temporal event streams

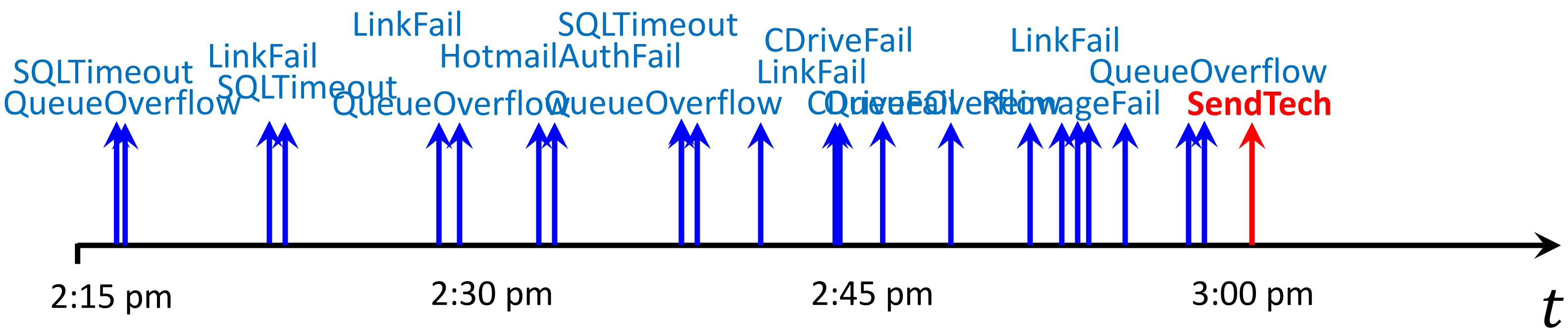
A temporal event stream is
a time-stamped stream of labeled events.



This type of data is pervasive:
datacenter event logs, search queries, …

Want to model:
What events will happen when,
based on what events have happened when.

Temporal Event Sequences: Event Logs from a Datacenter



$\mathcal{L}$ is the set of possible events (i.e., things that can happen)

$\mathcal{D} = \{(t_1, l_1), (t_2, l_2), (t_3, l_3), \dots (t_n, l_n)\}$ where $l \in \mathcal{L}$ and $t_i < t_{i+1}$

Marked point processes

Treat data as a realization of a marked point process:

$$x = (t_1, l_1), \dots, (t_n, l_n)$$

Forward in time likelihood:

$$p(x) = \prod_{i=1}^{n} p(t_i, l_i | h_i)$$

where the history $h_i = h_i(x) = (t_1, l_1), \dots, (t_{i-1}, l_{i-1})$

Any $p(t_i, l_i | h_i)$ can be represented via conditional intensities $\lambda_l(t_i | h_i)$:

$$p(t_i, l_i | h_i) = \prod_{l} \lambda_l(t_i | h_i)^{\mathbb{I}(l = l_i)} e^{-\int_0^{t_i} \lambda_l(\tau | h_i) d\tau}$$

Proof sketch

Given pdf $p(t)$ define:

$$\lambda(t) \triangleq \frac{p(t)}{1 - \underbrace{\int_0^t p(\tau)d\tau}_{P(t)}}$$

Then,

$$P'(t) = \lambda(t)[1 - P(t)]$$

Calculus:

$$P(t) = 1 - e^{-\int_0^t \lambda(\tau)d\tau}$$

$$p(t) = \lambda(t)e^{-\int_0^t \lambda(\tau)d\tau}$$

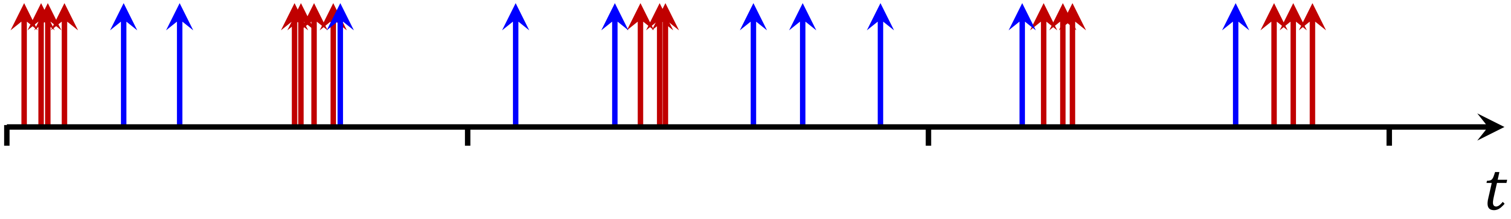
Given $p(t, l) = p(t)p(l|t)$ define

$$\lambda_l(t) \triangleq \lambda(t)p(l|t)$$

Then

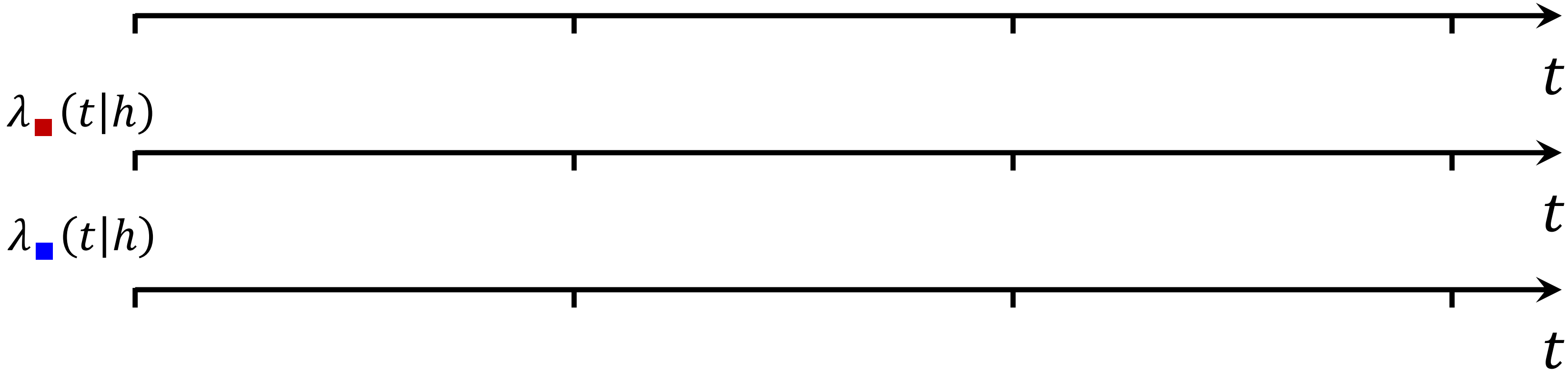
$$p(t, l') = p(t)p(l'|t) = \lambda_{l'}(t)e^{-\sum_l \int_0^t \lambda_l(\tau)d\tau}$$

Conditional intensities



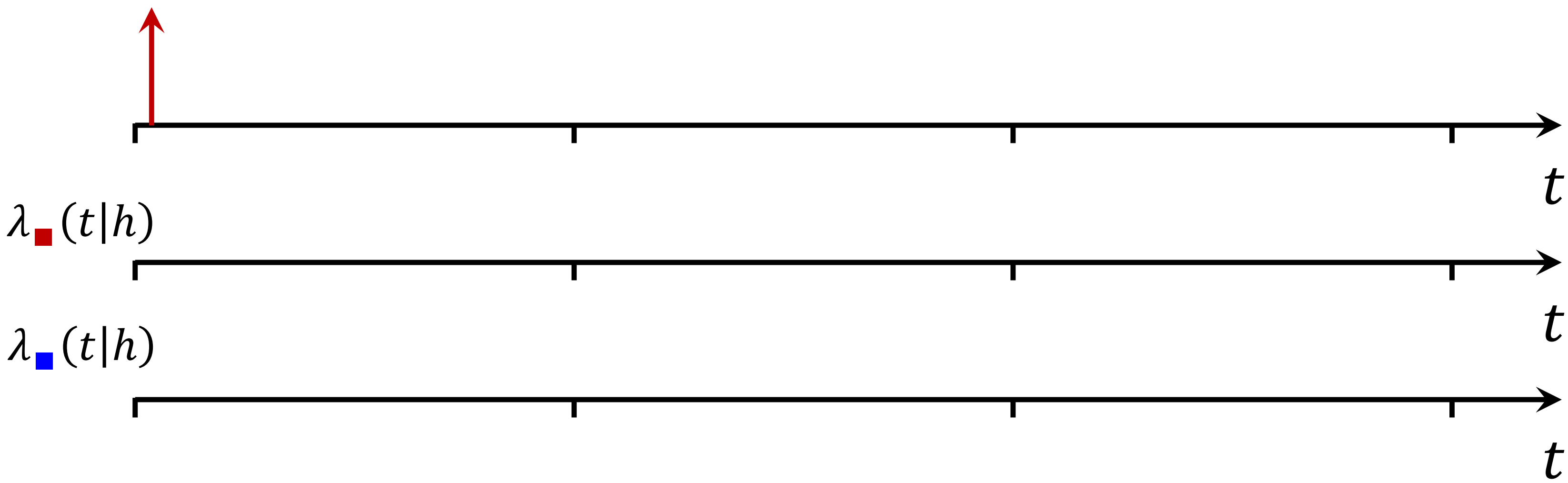
$$p(t_i, l_i | h_i) = \prod_l \lambda_l(t_i | h_i)^{\mathbb{1}(l = l_i)} e^{-\int_0^{t_i} \lambda_l(\tau | h_i) d\tau}$$

Conditional intensities



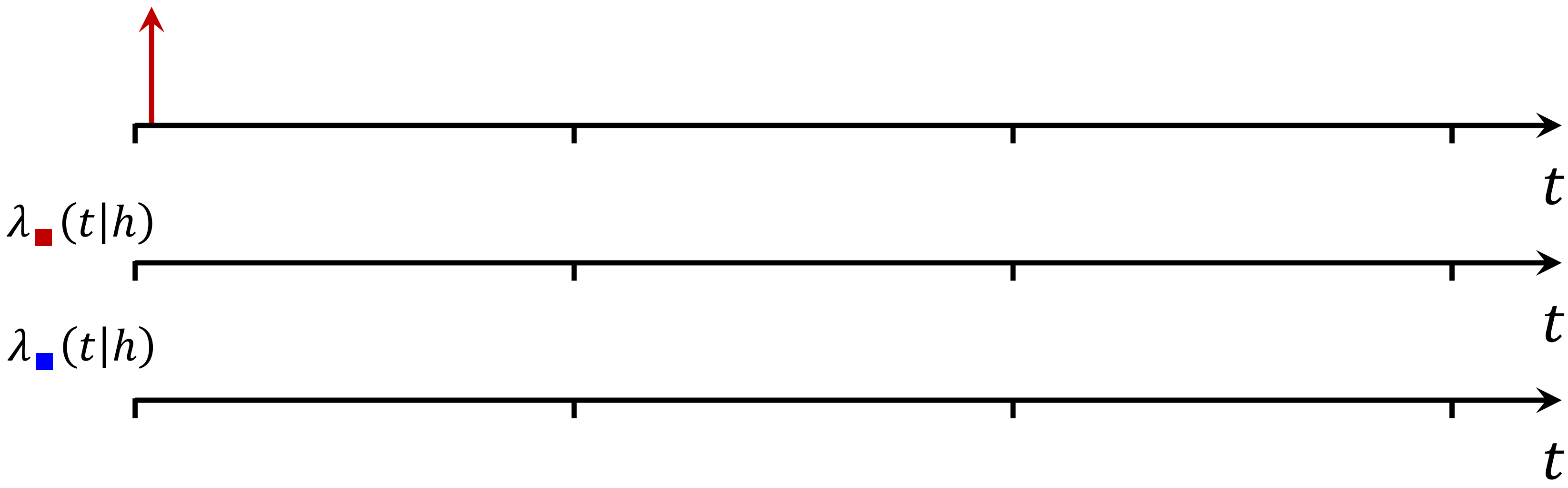
$$p(t_i, l_i | h_i) = \prod_{l} \lambda_l(t_i|h_i)^{\mathbb{1}(l=l_i)} e^{-\int_0^{t_i} \lambda_l(\tau|h_i) d\tau}$$

Conditional intensities



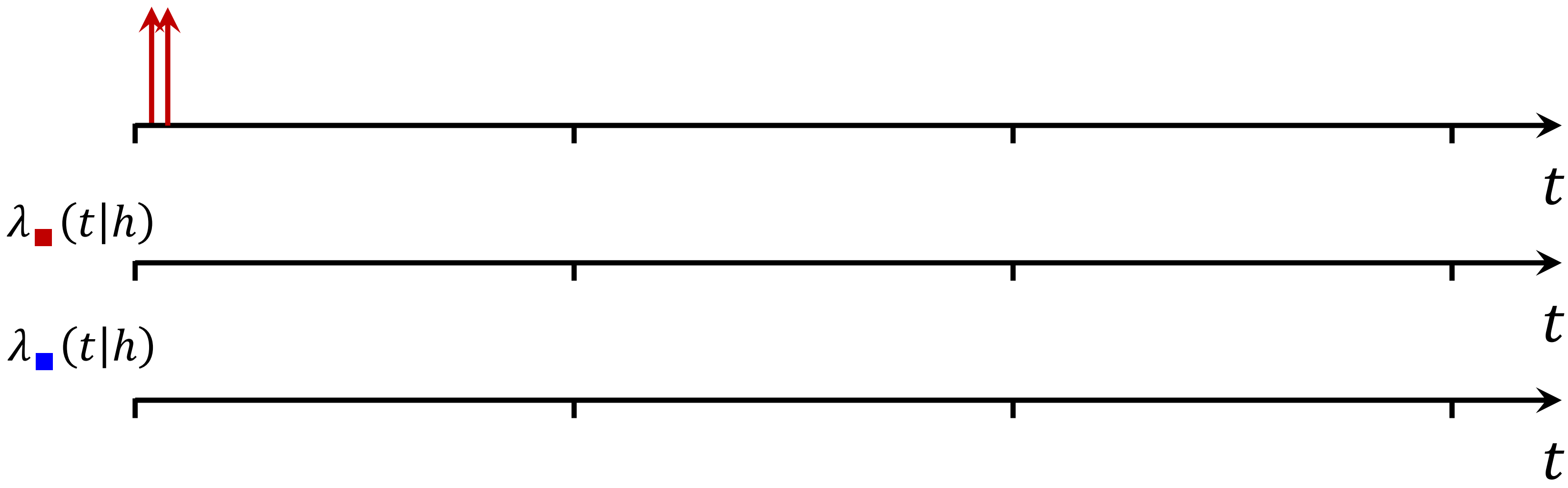
$$p(t_i, l_i | h_i) = \prod_l \lambda_l(t_i | h_i)^{\mathbb{1}(l = l_i)} e^{-\int_0^{t_i} \lambda_l(\tau | h_i) d\tau}$$

Conditional intensities



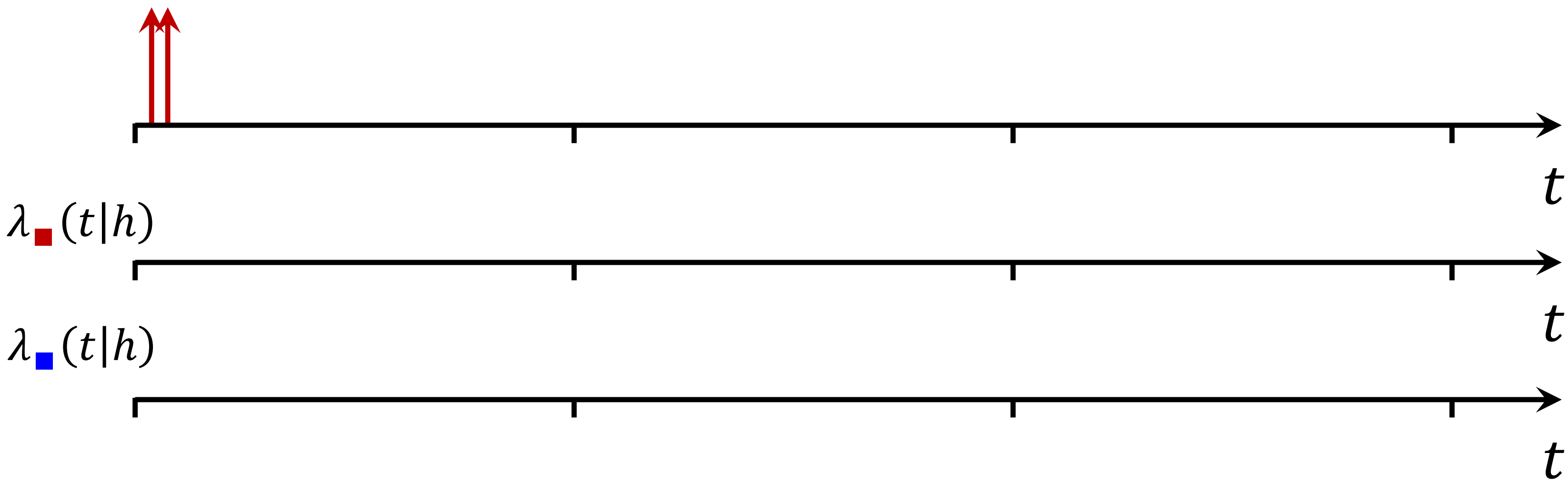
$$p(t_i, l_i | h_i) = \prod_l \lambda_l (t_i | h_i)^{\mathbb{1}(l = l_i)} e^{-\int_0^{t_i} \lambda_l (\tau | h_i) d\tau}$$

Conditional intensities



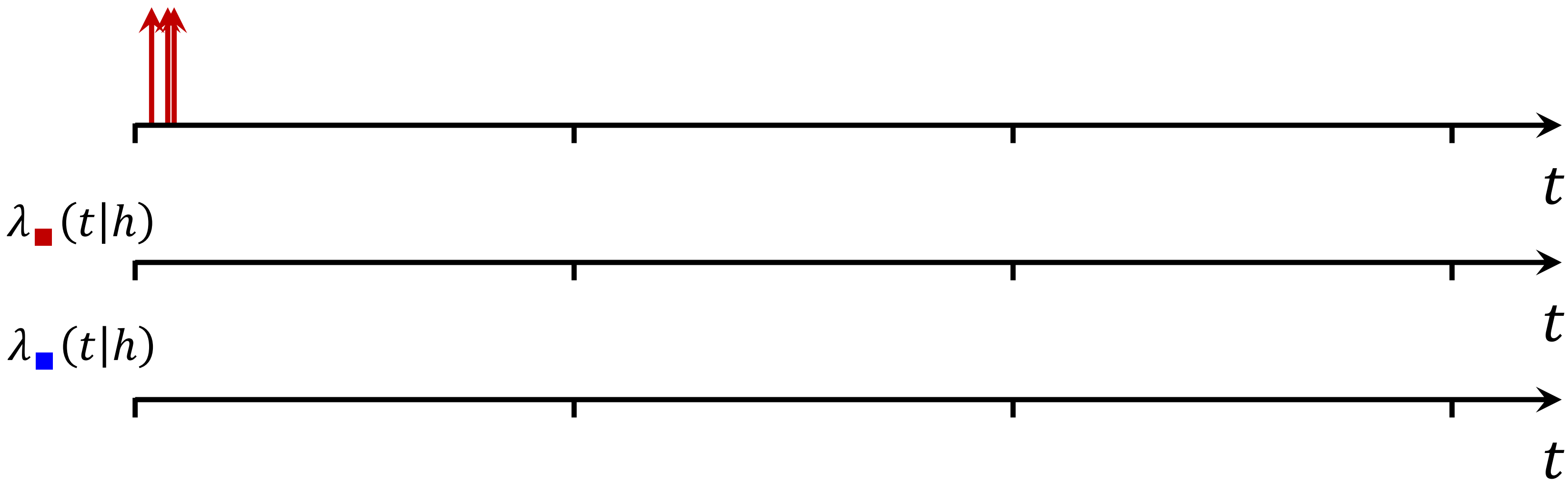
$$p(t_i, l_i | h_i) = \prod_l \lambda_l(t_i|h_i)^{\mathbb{1}(l=l_i)} e^{-\int_0^{t_i} \lambda_l(\tau|h_i) d\tau}$$

Conditional intensities



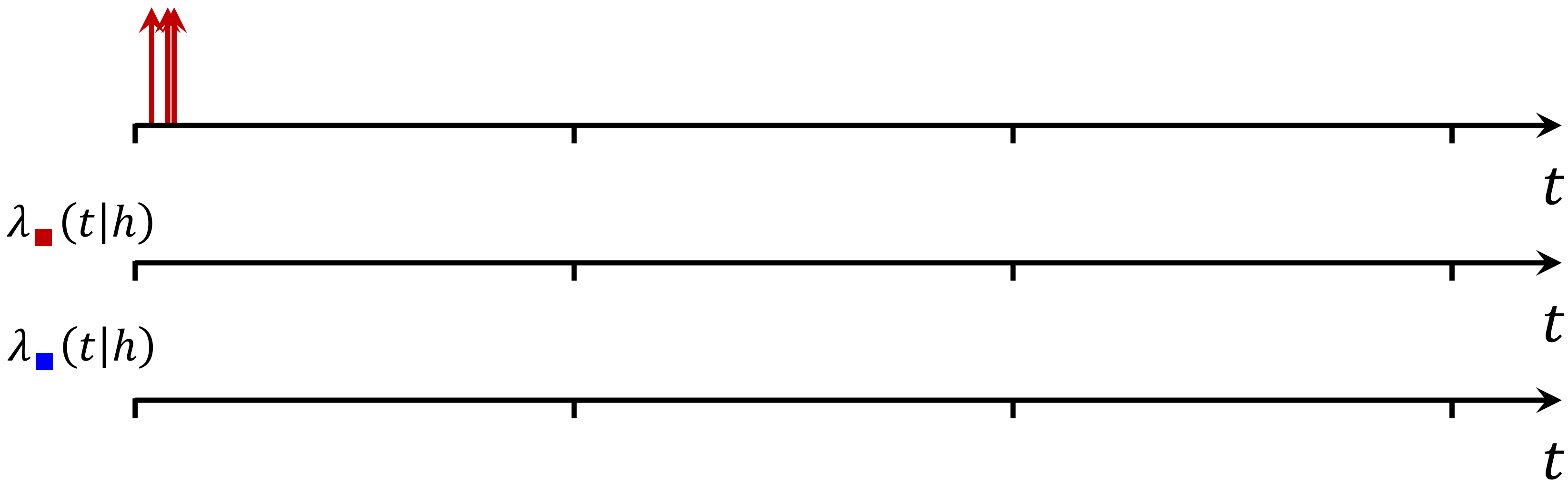
$$p(t_i, l_i | h_i) = \prod_l \lambda_l(t_i | h_i)^{\mathbb{1}(l = l_i)} e^{-\int_0^{t_i} \lambda_l(\tau | h_i) d\tau}$$

Conditional intensities



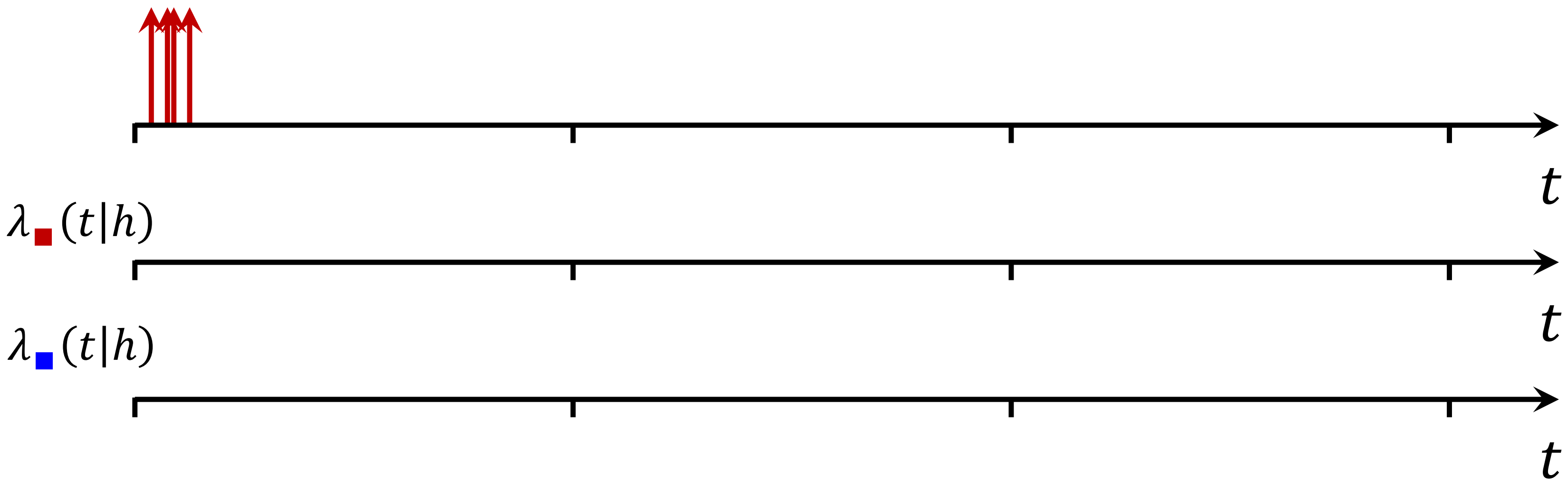
$$p(t_i, l_i | h_i) = \prod_l \lambda_l(t_i | h_i)^{\mathbb{1}(l = l_i)} e^{-\int_0^{t_i} \lambda_l(\tau | h_i) d\tau}$$

Conditional intensities



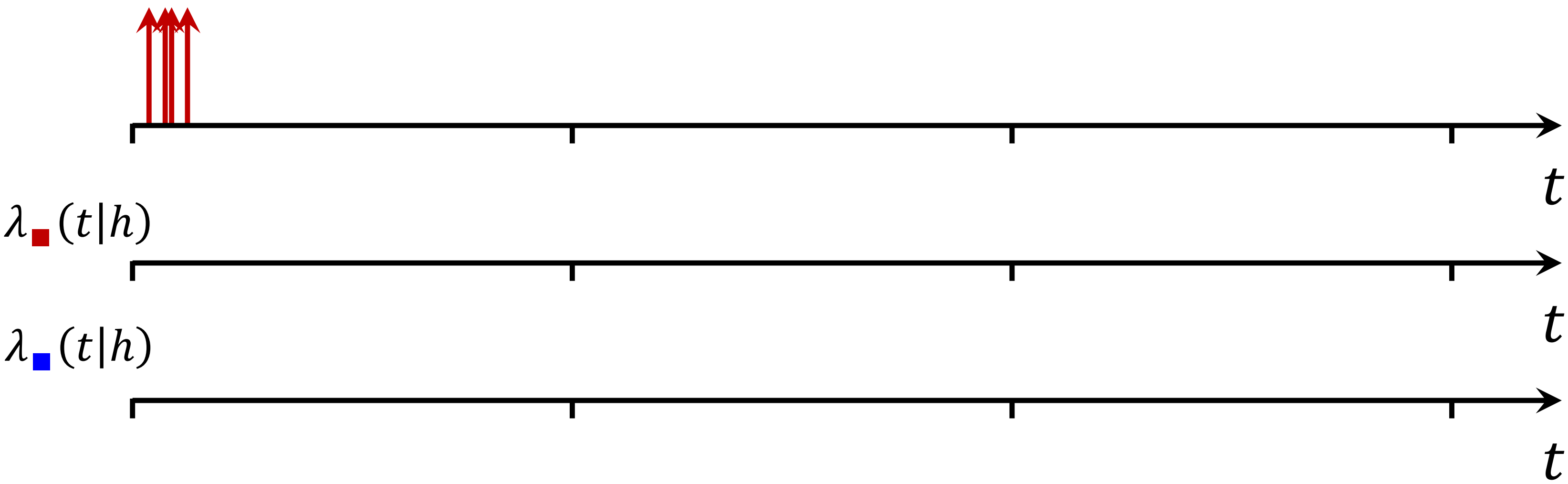
$$p(t_i, l_i | h_i) = \prod_l \lambda_l(t_i | h_i)^{\mathbb{1}(l = l_i)} e^{-\int_0^{t_i} \lambda_l(\tau | h_i) d\tau}$$

Conditional intensities



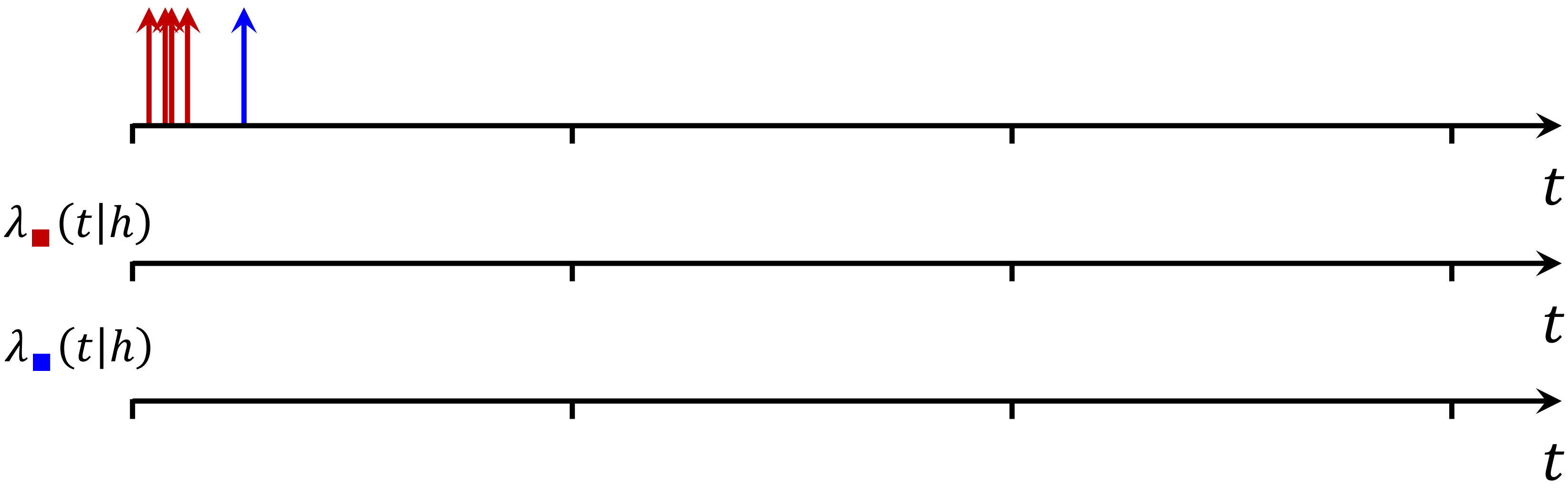
$$p(t_i, l_i|h_i) = \prod_{l} \lambda_l(t_i|h_i)^{\mathbb{1}(l=l_i)} e^{-\int_0^{t_i} \lambda_l(\tau|h_i)d\tau}$$

Conditional intensities



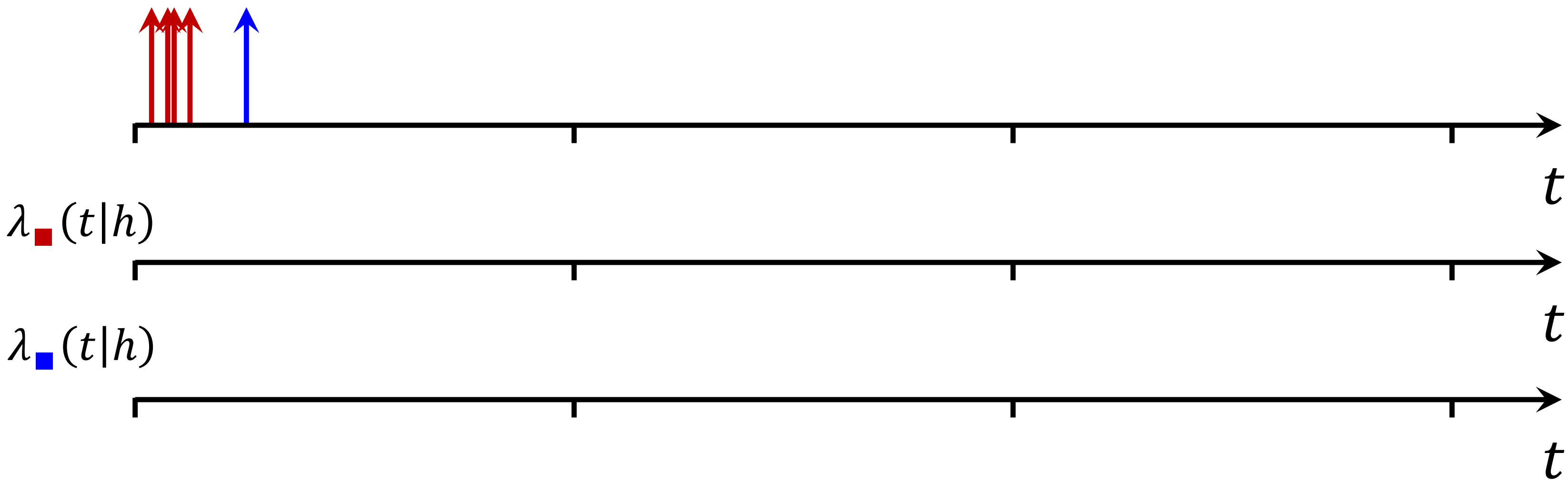
$$p(t_i, l_i | h_i) = \prod_l \lambda_l(t_i | h_i)^{\mathbb{1}(l = l_i)} e^{-\int_0^{t_i} \lambda_l(\tau | h_i) d\tau}$$

Conditional intensities



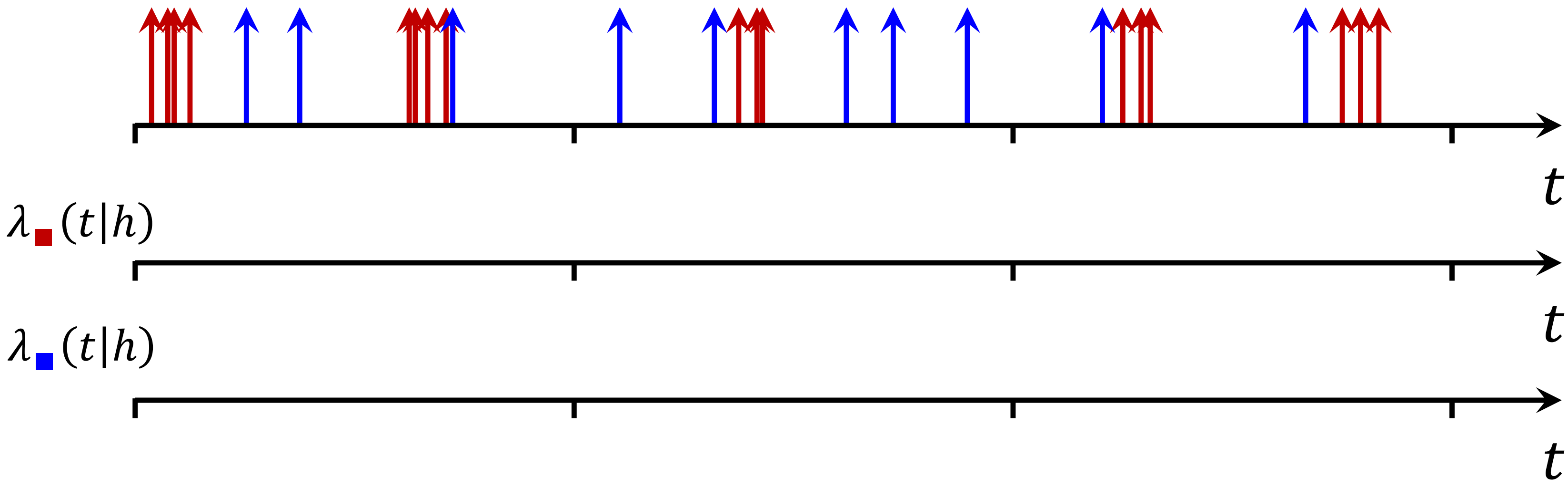
$$p(t_i, l_i | h_i) = \prod_l \lambda_l(t_i | h_i)^{\mathbb{1}(l = l_i)} e^{-\int_0^{t_i} \lambda_l(\tau | h_i) d\tau}$$

Conditional intensities



$$p(t_i, l_i | h_i) = \prod_l \lambda_l(t_i | h_i)^{\mathbb{1}(l = l_i)} e^{-\int_0^{t_i} \lambda_l(\tau | h_i) d\tau}$$

Conditional intensities



1) Behavior of different event types specified separately
2) Highlights dependency of each event types on history

Event Sequence Notation

Example: $\mathcal{L} = \{a, b, c\}$

$$\mathcal{D} = \{(t_1, l_1), \dots (t_n, l_n)\}$$

e.g., $\{(1, a\ ), (3, b), (t_3 = 5, a\ ), (8, c)\}$

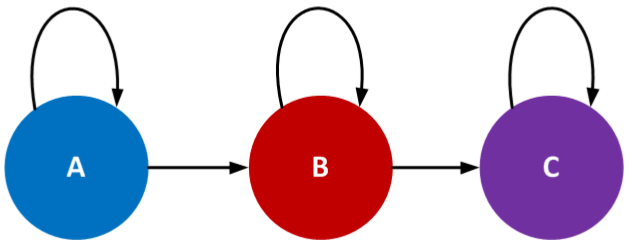
$h_i = h(t, \mathcal{D})$ is the ***history up to i***

$$h_3 = \{(1, a), (3, b), (5, a)\}$$

$[h]_A$ is the ***filtered history*** for $A \subseteq \mathcal{L}$

$$[\mathcal{D}]_a = \{(1, a), (5, a)\}$$

Graphical Event Models



A Graphical Event Model (GEM) is a pair $\langle G, \Theta \rangle$

Vertices for each event type $\mathcal{L} = \{a, b, c\}$

Edges represent potential dependencies

$\Theta_l \in \Theta$ parameterizes intensity function for l

$$\lambda_a(t|h, \Theta_a) = \lambda_a\left(t\middle|[h]_{\pi_a}, \Theta_a\right)$$

$$\lambda_b(t|h, \Theta_b) = \lambda_b\left(t\middle|[h]_{\pi_b}, \Theta_b\right)$$

$$\lambda_c(t|h, \Theta_c) = \lambda_c\left(t\middle|[h]_{\pi_c}, \Theta_c\right)$$

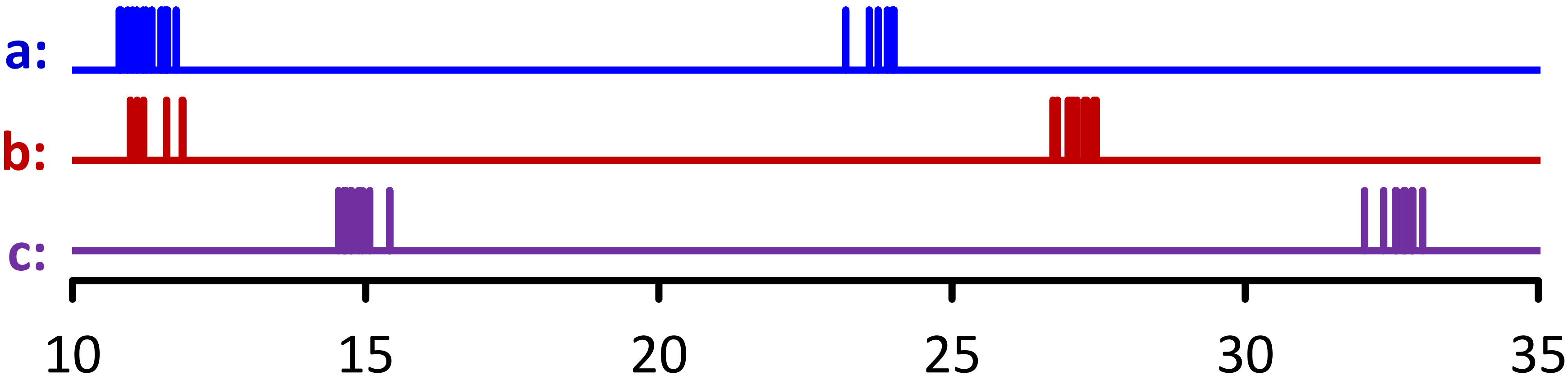
Learning Graphical Event Models

- Specify functional form(s) for intensities λ_l with separate parameters for each event
- Likelihood factors according to $\mathcal{L}$ so we can learn each intensity function separately.
  - Bayesians also require factorization of prior
- Search over space of directed graphs
  - Add/remove parents that improve the score

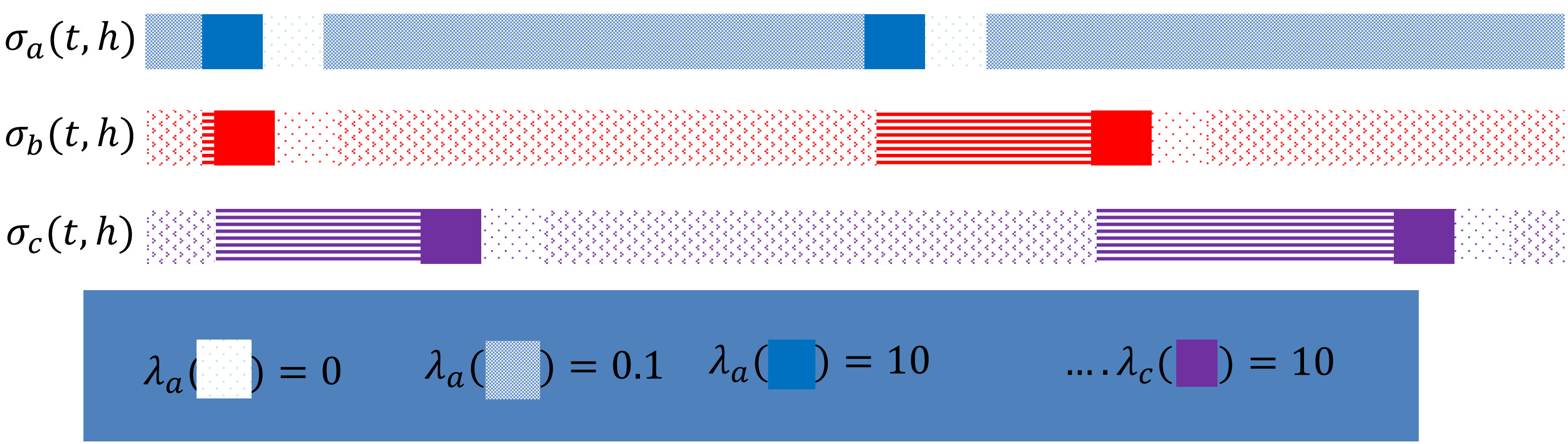
Piecewise-Constant CIMs (PCIM)

- Idea: restrict $\lambda_l(t_i|h_i)$ to be piecewise constant in t for all event sequences
- A state function $\sigma(t, h)$ maps histories to a discrete set of states Σ
- A PCIM is a pair $\mathcal{M} = \langle S, \Theta \rangle$ where
  - Structure S= $\{\langle \sigma_l(t, h), \Sigma_l \rangle\}_{l \in \mathcal{L}}$
  - Parameters $\Theta = \{\Theta_l\}_{l \in \mathcal{L}}$ and $\Theta_l = \{\lambda_{ls}\}_{s \in \Sigma_l}$

Piecewise-Constant CIMs



State Functions (coloring)



Piecewise-Constant CIM

- A PCIM (CIM) $\mathcal{M} = \langle S, \Theta \rangle$ (where $\Theta = \{\Theta_1, \ldots, \Theta_{|\mathcal{L}|}\}$ and $\Theta_l = \{\lambda_{ls}\}_{s \in \Sigma_l}$) has likelihood

$$p(\mathcal{D}|\mathcal{M}) = \prod_{l \in \mathcal{L}} \prod_{s \in \Sigma_l} \lambda_{ls}{}^{c(l,s)} \, e^{-\lambda_{ls} d(l,s)}$$

$c(l, s)$ is the count of event l when $\sigma_l(t, h) = s$ in $\mathcal{D}$

$d(l, s)$ is the total duration of $\sigma_l(t, h) = s$ in $\mathcal{D}$

Piecewise-Constant CIM

Product of Gammas is conjugate prior, even though the likelihood isn't a product of exponentials!

$$p(\lambda_{ls}|\alpha_{ls},\beta_{ls}) = \frac{\beta_{ls}}{\Gamma(\alpha_{ls})}\lambda_{ls}^{\alpha_{ls}-1}e^{-\beta_{ls}\lambda_{ls}}$$

Closed-form posterior:

$$p(\lambda_{ls}|\alpha_{ls},\beta_{ls},\mathcal{D},S) = p(\lambda_{ls}|\alpha_{ls}+c(l,s),\beta_{ls}+d(l,s)\,)$$

Closed-form marginal likelihood:

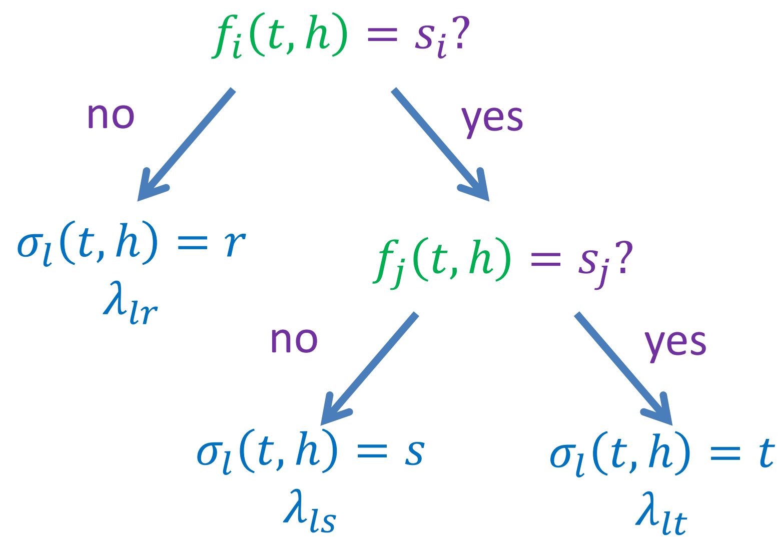
$$p(\mathcal{D}|S) = \prod_{ls}\gamma_{ls}(\mathcal{D}) \qquad \gamma_{ls}(\mathcal{D}) = \frac{\beta_{ls}^{\alpha_{ls}}}{\Gamma(\alpha_{ls})}\frac{\Gamma(\alpha_{ls}+c(l,s))}{(\beta_{ls}+d(l,s))^{\alpha_{ls}+c(l,s)}}$$

Piecewise-Constant CIM

Defining PCIM Structures

- Let $\mathcal{B} = \{f_1(t,h), \ldots, f_n(,)\}$ where $f_i(t,h)$ is a basis state function (BSF)
- A family of structures $\mathcal{S}(\mathcal{B})$ is obtained by combining BSFs. We use decision trees but one could use decision graphs.

Example

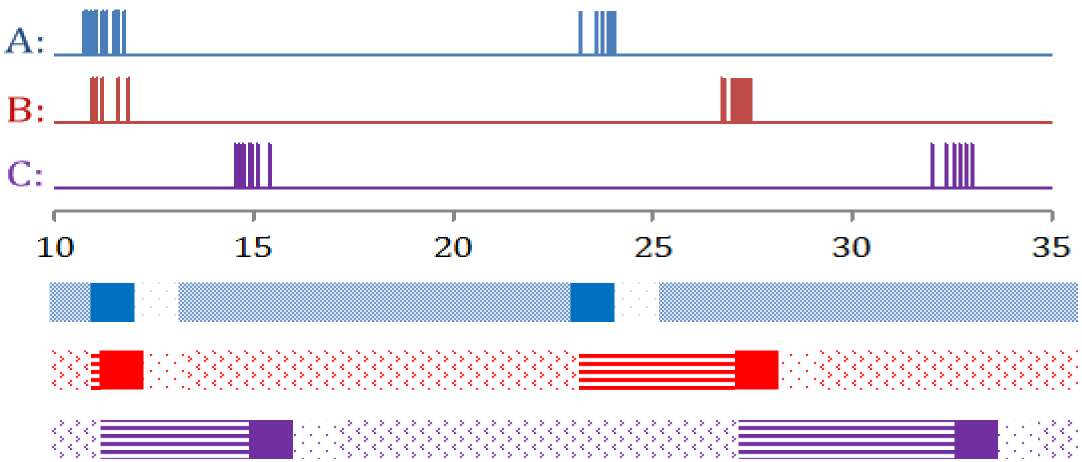


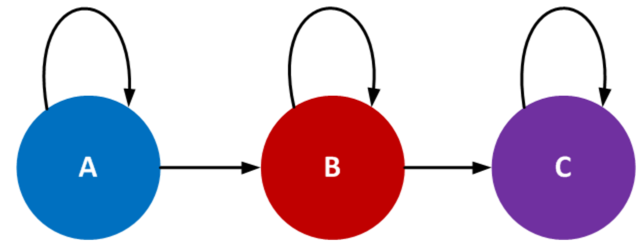
Piecewise-Constant CIM

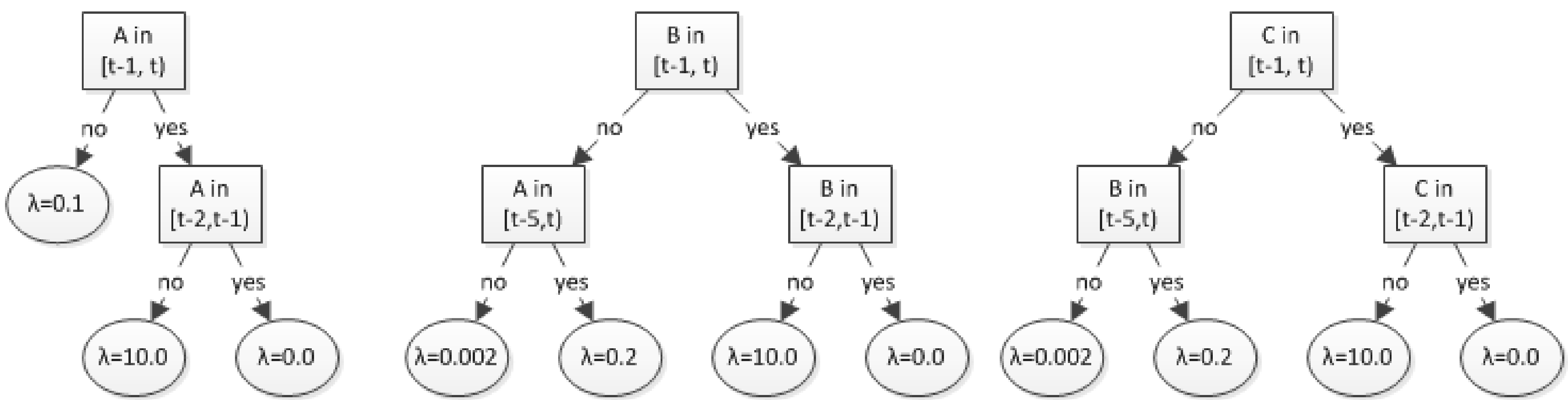
Example Types of basis state functions $f(t, h)$

- Event-type specific state functions
  - $f(t, h) = f(t, [h]_l)$ depends only on the history of a specific event type
- Windowed state functions
  - $f(t, h) = f\left(t, \{h\}_{(t-s, t-e)}\right)$ depends only on the history during a window relative to time t.
- Historical state functions
  - $f(t, h)$ depends on the "last" events that have happened but not their times.

Piecewise-Constant CIM







$\sigma_a(t,h)$ $\sigma_b(t,h)$ $\sigma_c(t,h)$

Piecewise-Constant CIM:Learning

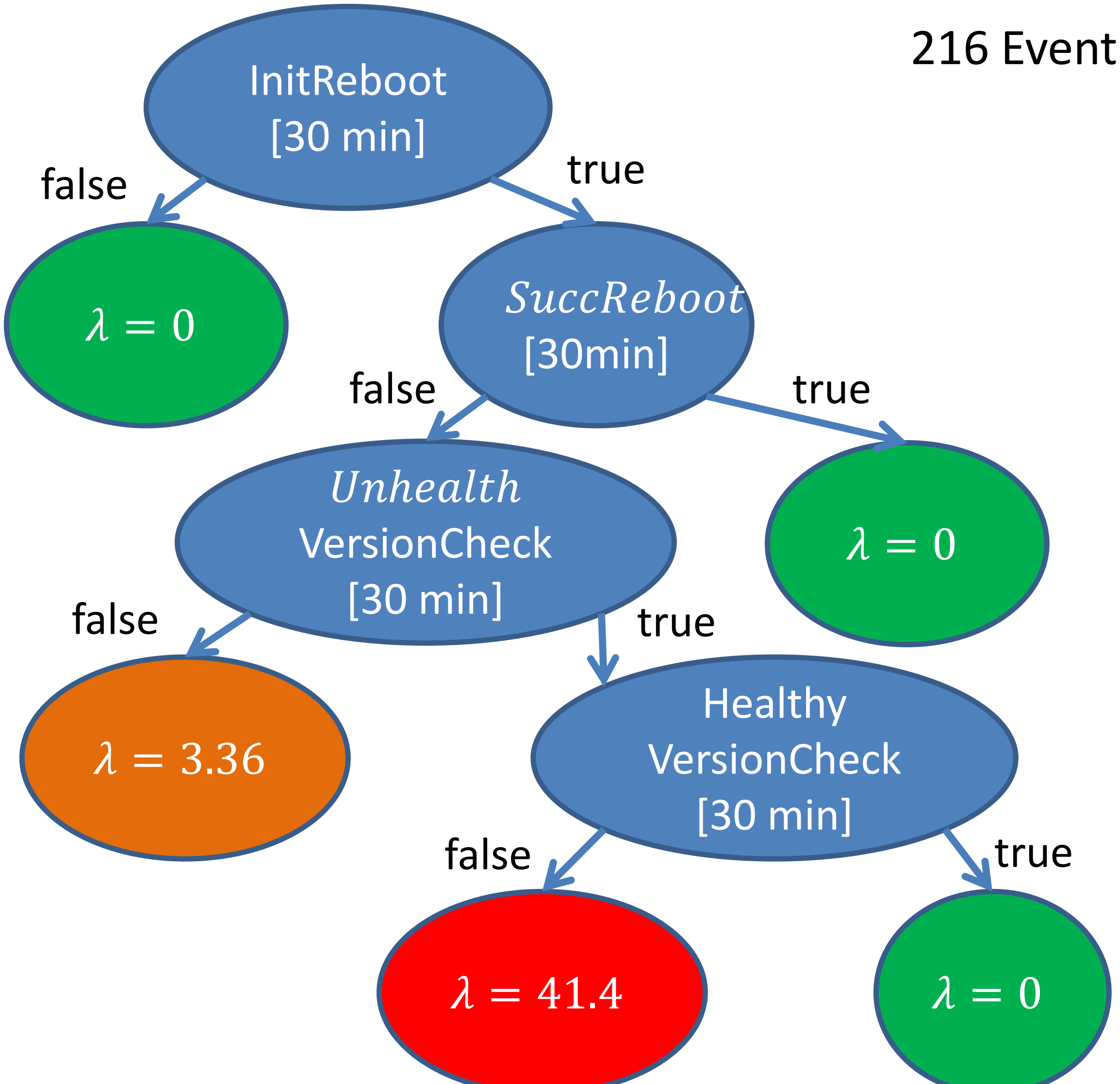
We use a Bayesian Model selection approach to choose $S \in \mathcal{S}(\mathcal{B})$

- For each $l \in \mathcal{L}$
  - Start with empty decision tree (i.e., $\forall t, \forall h\ \sigma_l(t,h) = k$)
  - For each leaf in decision tree
    - Evaluate each possible split ($f \in \mathcal{B}, s$)
    - Choose split that most improves the marginal likelihood

Alternatively one could use MCMC to average over $\mathcal{S}(\mathcal{B})$.

Example: $\lambda_{RebootFail}$ decision tree

216 Event types



Learning Causal Graphical Models

Directed Acyclic Assumption (DAG) causal model can be represented by a directed acyclic graph $G = \langle X, E \rangle$

Data Assumption: world is described by some $P(X)$ and the observed world is some $P(O)\ O \subseteq X$

Reliable Information Assumption (Reliable) the world provides reliable information about independencies among observed variables $\mathcal{O} \subseteq X$.

- $I(A, C, B)$ means A is independent of B given C

Learning Causal Graphical Models

Assumptions that connect observed world P and causal model G

Causal Markov Assumption (CMA):

$$\text{If } d_G(A,B,C) \Rightarrow I_P(A,B,C)$$

Note 1: $d_G(A,B,C)$ is d-separation: a vertex separation criterion
Note 2: A graphical model "non-causal" Markov w.r.t. $P(X)$ it defines

Causal Faithfulness Assumption (CFA):

$$\text{If } I_P(A,B,C) \Rightarrow d_G(A,B,C)$$

Learning Causal Graphical Models

Learning Scenarios

- *Complete data*: All variables observed ($X = \mathcal{O}$)
- *Causal sufficiency*: No pair of observed variables have unobserved common ancestors.
- *General case*: $\mathcal{O} \subseteq X$

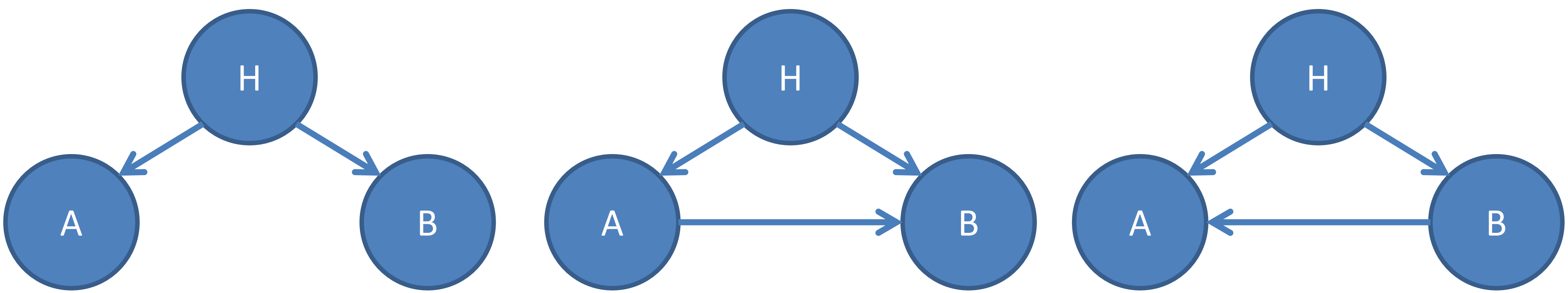
Goal: In each learning scenario, use independence facts to identify causal information common to every graph with those independencies/separation facts.

Learning Causal Graphical Models

Mantra: “Correlation does not imply causation”

Complete data: Indistinguishable

General case: Also indistinguishable



Learning Causal Graphical Models

Interesting general case result

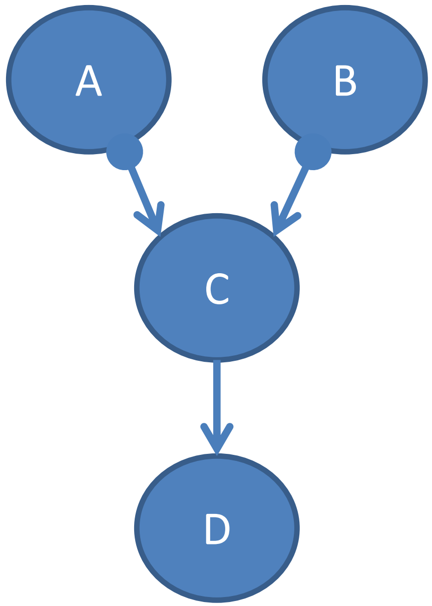
Under the assumptions

DAG, Reliable, CMA and CFA

If the only independence facts we observe to hold are

$I(A, \emptyset, B), I(A, C, D), I(B, C, D)$

then C is a cause of D.



Learning Causal GEMs

Step 1: Change the separation criterion from d-separation to δ-separation

$$\delta(\mathrm{A}, \mathrm{C}, \mathrm{B}) \text{ in } G = \langle \mathcal{L}, \mathcal{E} \rangle \text{ if and only if } d(A, C, B) \text{ in } G^B$$
$$\text{where } G^B = \langle \mathcal{L}, \mathcal{E}^B \rangle \text{ and } \mathcal{E}^B = \{\langle l_1, l_2 \rangle \in \mathcal{E} | l_1 \notin B\}$$

Step 2: Change from independence tests to factorization (process independence) tests

Step 3: Assume analog of CMA, CFA, Reliable

Step 4: Prove things

Learning Causal GEMs

Learning Scenarios

- *Complete data*: All variables observed ($X = \mathcal{O}$)
  - Result: Can recover the structure.
- *Causal sufficiency*: No pair of observed variables have unobserved common ancestors.
  - Result: Can recover the structure over $\mathcal{O}$. (all causes)
- *General case*: $\mathcal{O} \subseteq X$
  - In Progress:
    - Some sufficient conditions for cause
    - Some sufficient conditions for non-cause
    - Some sufficient conditions for existence of unmeasured common cause

Open Issues

- Characterize what one can learn in the general case
- Justification of using Process Independence to learn cause (e.g., completeness of δ-separation)
- Principled approaches to testing process independence statements.
- Consistency of score learning for process independence.
- Relaxing assumptions such as the reliability assumption