



Universidad de Granada

**decsai.ugr.es**

## Minería de Procesos

**Juan Fernández (teoría), Luis Castillo (prácticas)**



**Departamento de Ciencias de la  
Computación e Inteligencia Artificial**

Descubrimiento de modelos de proceso, que describen el comportamiento de personas/agentes físicos/agentes virtuales, a partir del registro de su traza de actividad.

- Sesión 1: Process Mining.
  - conocer cómo se aprenden modelos de proceso a partir de un log, en el que se registran la traza de actividad de personas/procesos.
  - Sesión Práctica 1.
- Sesión 2: Planificación automática
  - conocer cómo representar modelos de proceso que describen el **comportamiento orientado a objetivos** de agentes/sistemas autónomos, mediante las técnicas de planificación automática de tareas
  - Sesión Práctica 2.
- Sesión 3: Aprendizaje en planificación
  - conocer cómo **aprender dominios de planificación automática** a partir de la traza de planes.
  - Sesión Práctica 3



Universidad de Granada

**decsai.ugr.es**

# Process Mining

**Juan Fernández Olivares**



**Departamento de Ciencias de la  
Computación e Inteligencia Artificial**

- 1. Motivación (¿por qué necesitamos PM?)**
- 2. Introducción (¿qué es PM?)**
- 3. Ciclo de vida en BPM (Business Process Management)**
- 4. Visión general (cómo encaja Process Mining en el mundo) ( PM Book: Cap.1)**
- 5. Logs de Eventos**
- 6. Modelos de proceso (PM Book: Cap.2 y 3.)**
- 7. Relaciones entre logs y modelos de proceso**
- 8. Una notación para modelos de proceso: Redes de Petri (PM Book: Cap. 3)**
- 9. WF-nets: un tipo especial de Redes de Petri para aprender modelos de proceso(PM Book: Cap. 3)**
- 10. Aprendizaje de modelos de proceso: Process Discovery (PM Book: Cap.6)**
  - 1. Algoritmo alpha**
  - 2. Limitaciones del algoritmo alpha.**
  - 3. Criterios de Calidad de un algoritmo de Process Discovery**
- 11. Técnicas Avanzadas (PM Book: Cap 7)**
  - 1. Process Mining Heurístico**
  - 2. Inductive Mining**

- Van der Aalst, Wil M. P. *Process Mining*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2016.
  - Acceso e-book desde VPN UGR:<http://link.springer.com/book/10.1007%2F978-3-662-49851-4>
- Aalst, Van Der, y Wil M. P. «Business Process Management: A Comprehensive Survey». *International Scholarly Research Notices* 2013 (12 de febrero de 2013): e507984. doi:10.1155/2013/507984.
  - Acceso VN UGR: <http://www.hindawi.com/journals/isrn/2013/507984/>
- Ferreira, Diogo R. *A Primer on Process Mining: Practical Skills with Python and Graphviz*. New York: Springer, 2017.
  - Acceso e-book desde VPN UGR:  
<https://link.springer.com/openurl?genre=book&isbn=978-3-319-56426-5>.
- Material muy interesante:
  - <http://www.processmining.org/>. Libros, artículos, presentaciones, herramientas para Process Mining.
  - <http://www.workflowpatterns.com/>. Características comunes que debe reunir cualquier modelo de proceso.

- Data Mining parte de conjuntos de datos que constan de instancias de individuos, entidades, objetos, ....
- Variables: atributos o propiedades de las instancias
  - Numéricas, ordinales (alto, medio, bajo) o nominales (sí/no, rojo-verde-azul)
- Para responder a preguntas buscando relaciones entre los datos.

## Data set 1

Data about 860 recently deceased persons to study the effects of drinking, smoking, and body weight on the life expectancy.

drinker	smoker	weight	age
yes	yes	120	44
no	no	70	96
yes	no	72	88
yes	yes	55	52
no	yes	94	56

### Questions:

- What is the effect of smoking and drinking on a person's bodyweight?
- Do people that smoke also drink?
- What factors influence a person's life expectancy the most?
- Can one identify groups of people having a similar lifestyle?

©Wil van der Aalst & TU/e (use only with permission & acknowledgements)

## Data set 2

Data about 420 students to investigate relationships among course grades and the student's overall performance in the Bachelor program.

linear algebra	logic	program-ming	operations research	workflow systems	...	duration	result
9	8	8	9	9	...	36	cum laude
7	6	-	8	8	...	42	passed
-	-	5	4	6	...	54	failed
8	6	6	6	5	...	38	passed

### Questions:

- Are the marks of certain courses highly correlated?
- Which electives do excellent students (cum laude) take?
- Which courses significantly delay the moment of graduation?
- Why do students drop out?
- Can one identify groups of students having a similar study behavior?

©Wil van der Aalst & TU/e (use only with permission & acknowledgements)

En Data Mining pretendemos explicar o descubrir relaciones:

- Aprendizaje supervisado: explicar una variable de respuesta a partir de variables predictoras.
  - Clasificación y regresión.
- Aprendizaje no supervisado: las variables no tienen etiquetas predefinidas
  - Clustering (k-medias, ...)
  - Pattern Discovery (reglas de asociación).

## Data set 1

Data about 860 recently deceased persons to study the effects of drinking, smoking, and body weight on the life expectancy.

drinker	smoker	weight	age
yes	yes	120	44
no	no	70	96
yes	no	72	88
yes	yes	55	52
no	yes	94	56

### Questions:

- What is the effect of smoking and drinking on a person's bodyweight?
- Do people that smoke also drink?
- What factors influence a person's life expectancy the most?
- Can one identify groups of people having a similar lifestyle?

©Wil van der Aalst & TU/e (use only with permission & acknowledgements)

## Data set 2

Data about 420 students to investigate relationships among course grades and the student's overall performance in the Bachelor program.

linear algebra	logic	program-ming	operations research	workflow systems	...	duration	result
9	8	8	9	9	...	36	cum laude
7	6	-	8	8	...	42	passed
-	-	5	4	6	...	54	failed
8	6	6	6	5	...	38	passed

### Questions:

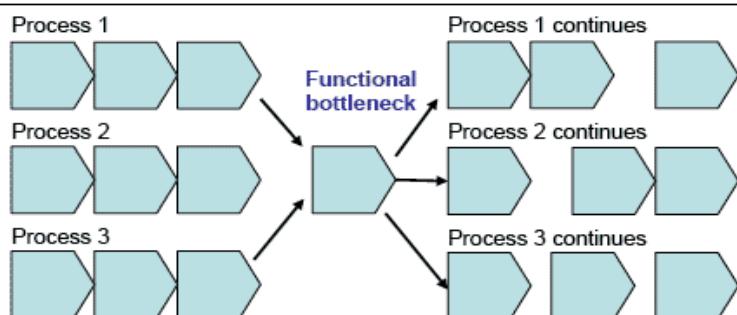
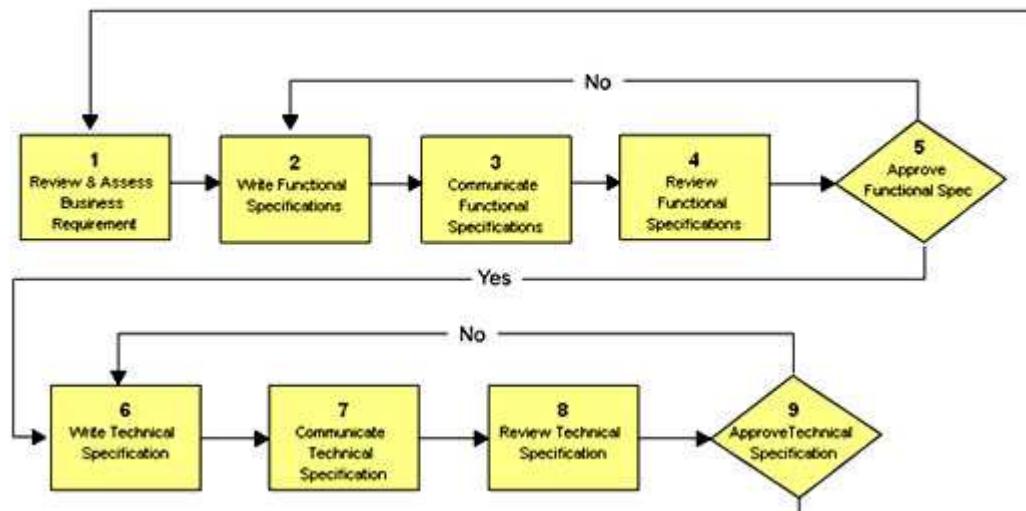
- Are the marks of certain courses highly correlated?
- Which electives do excellent students (cum laude) take?
- Which courses significantly delay the moment of graduation?
- Why do students drop out?
- Can one identify groups of students having a similar study behavior?

©Wil van der Aalst & TU/e (use only with permission & acknowledgements)

- Estamos interesados en el **registro de la actividad** de personas en una organización (hospital, banco, un curso e-learning, una oficina de proyectos, jugadores videojuegos....)
- Queremos responder a preguntas que requieren **conocer el modelo de proceso** para buscar la respuesta.
- En general comportamiento de agentes en un entorno.



- ¿Cuál es el proceso que realmente sigue la gente?
  - ¿Dónde están los cuellos de botella en mi proceso?
  - ¿Dónde se desvía la gente (o las máquinas) del proceso que yo he idealizado?.



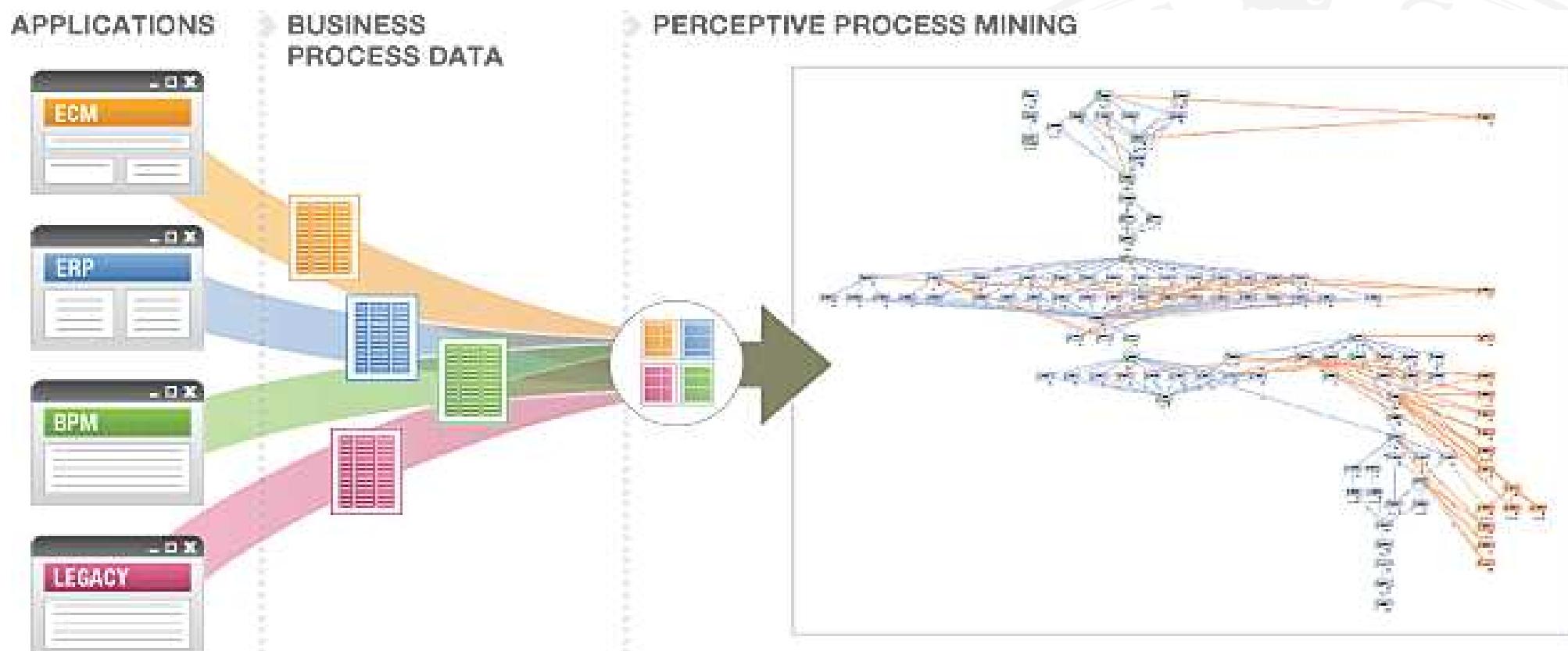
• Un conjunto de técnicas que relacionan

- **logs de eventos:**

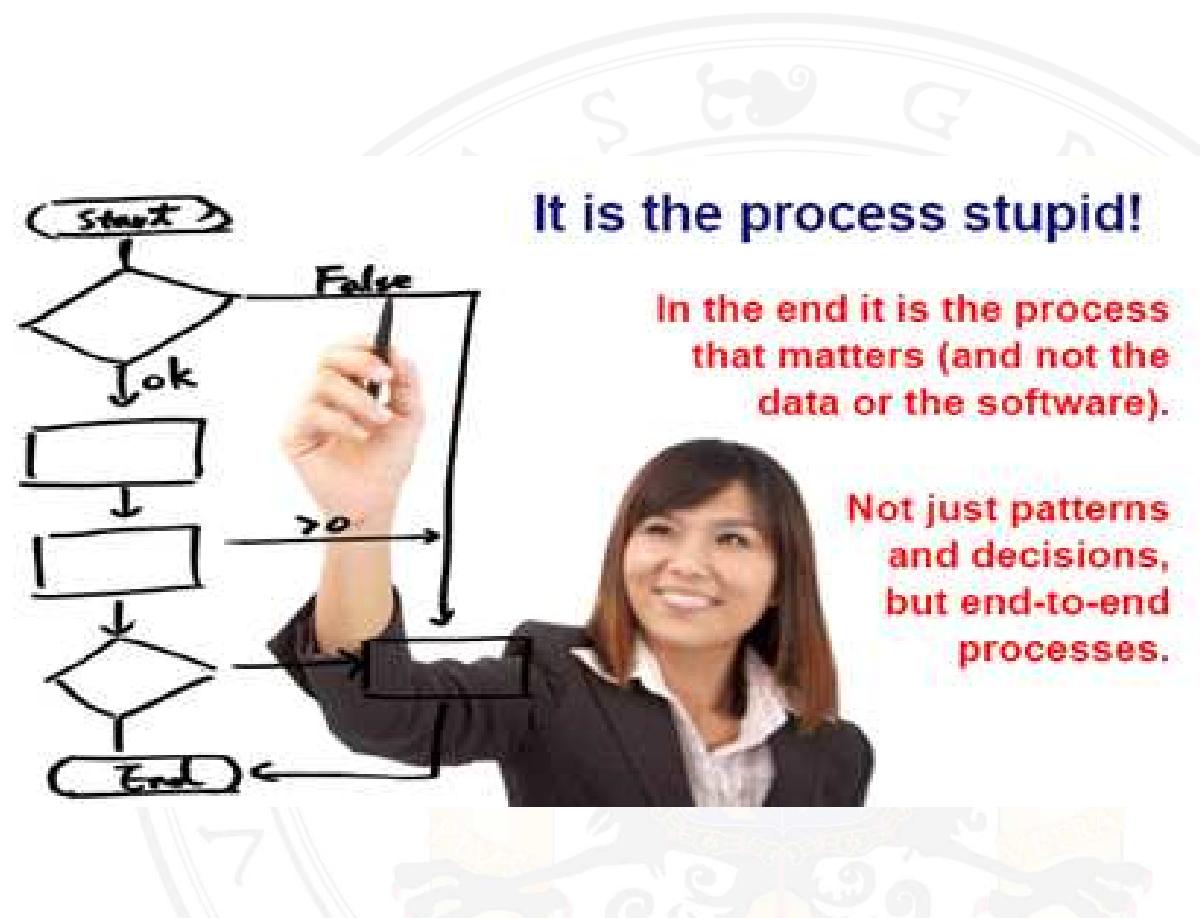
- la traza de actividades registrada por el comportamiento de uno o más agentes, almacenada en algún medio

- **modelos de proceso:**

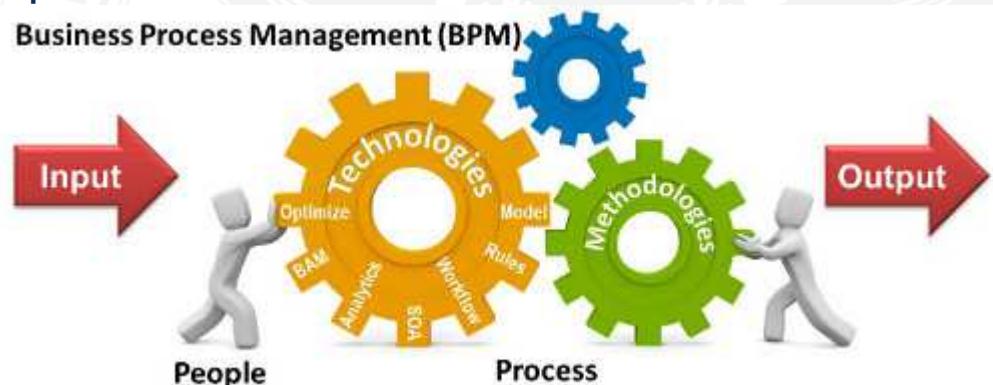
- especificación del flujo de control (secuencia, paralelo, condicional, ciclos, subprocessos) entre actividades, considerando recursos (humanos y materiales), tiempo, etc.



- PM ofrece respuesta a estas preguntas (distintas a las planteadas en DM)
  - Mediante el **descubrimiento** de procesos
  - Mediante la **comprobación de conformidad**: adecuación del modelo con las trazas de actividades registradas.
  - Mediante **la mejora de un modelo** de proceso ya existente.
- Usa técnicas próximas a Data Mining, pero van más allá de la mera detección de patrones o de decisiones simples y aisladas.
  - Interés en procesos “end-to-end”, se inician con una transacción de un usuario y finalizan devolviendo un resultado.



- Pueden usarse para:
  - Discutir/analizar responsabilidades en una organización.
  - Analizar el cumplimiento de los procedimientos de una organización
  - Predicciones sobre rendimiento usando simulación
  - Configurar los sistemas de una organización.
  - Recomendar actuaciones, basadas en el modelo de proceso
  - ...
- De todos estos aspectos relacionados con modelos de proceso se encarga BPM (Business Process Management):
  - Una disciplina que incorpora representación y gestión del conocimiento, tecnologías de la información y ciencias de la gestión para estudiar los aspectos operativos de los procesos que se llevan a cabo en una organización.

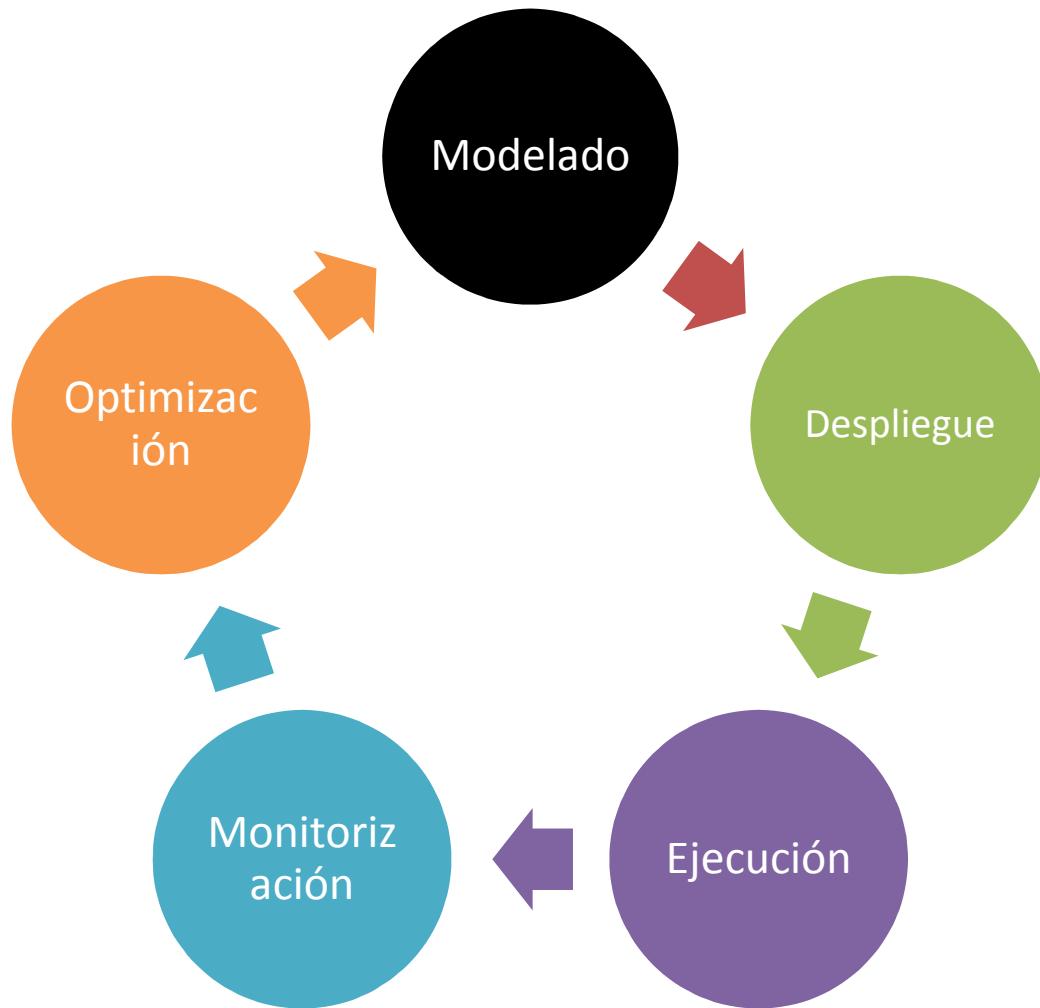


- BPM
  - Qué es
    - conjunto de técnicas para
      - *automatizar las tareas que se llevan a cabo en una organización (workflow), considerando recursos humanos y materiales*
    - Proceso de negocio,
      - *un conjunto de actividades que debe llevarse a cabo en una organización para cumplir una tarea o un objetivo.*
  - Ciclo de vida
    - Modelado, Instanciación, ejecución, monitorización y revisión

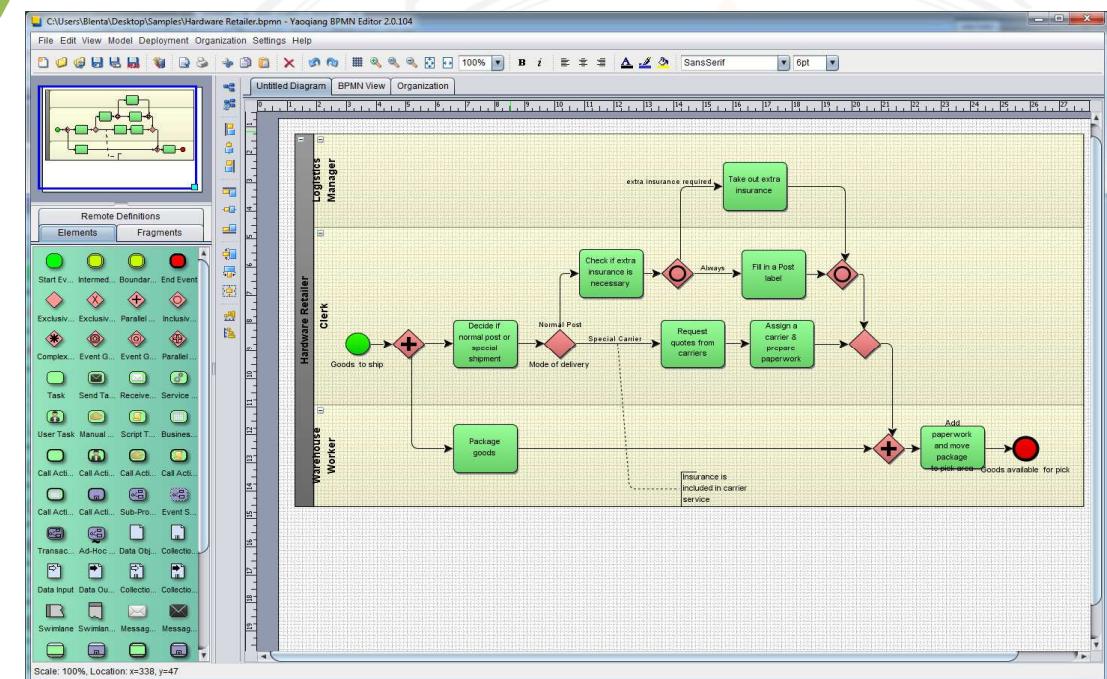


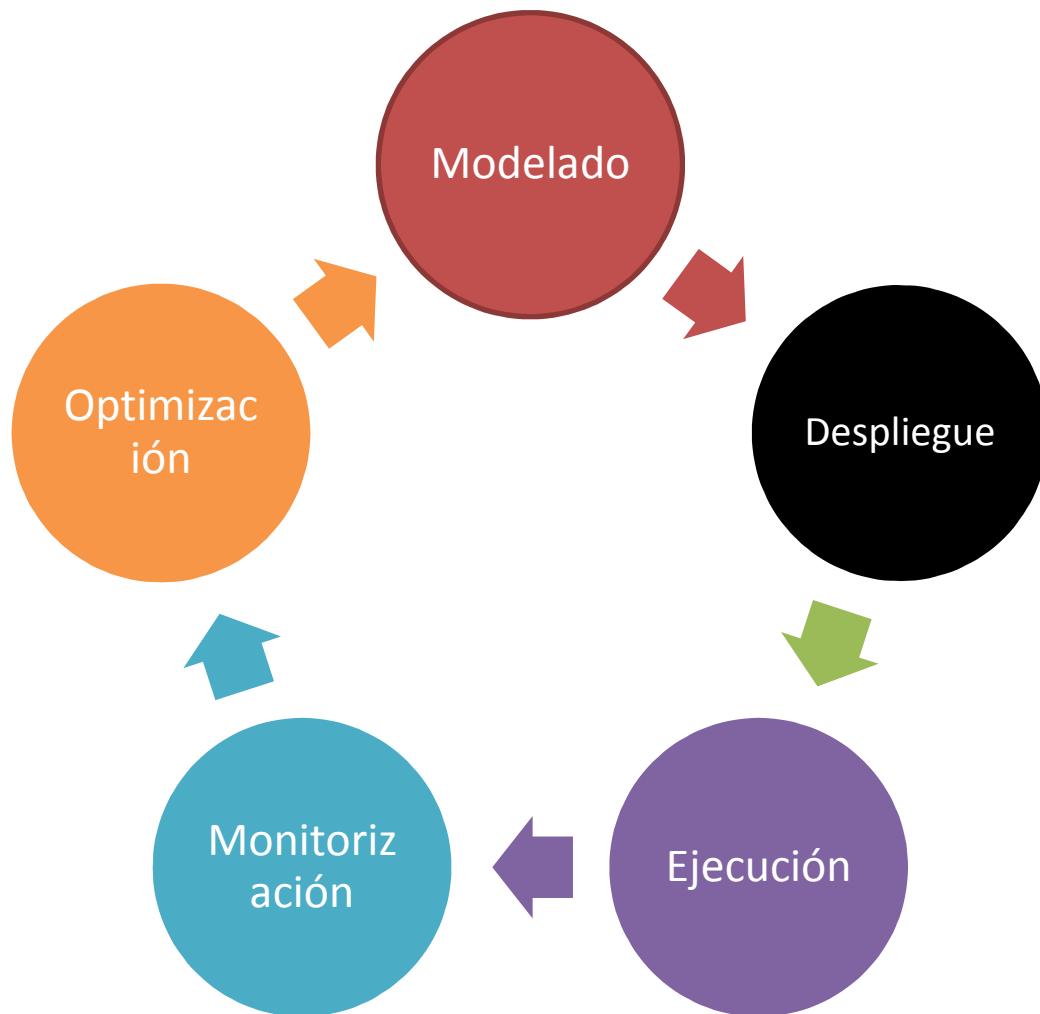
Aalst, Van Der, y Wil M. P. «Business Process Management: A Comprehensive Survey». *International Scholarly Research Notices* 2013 (12 de febrero de 2013): e507984. doi:10.1155/2013/507984.  
Acceso VPN UGR: <http://www.hindawi.com/journals/isrn/2013/507984/>

Weske, Mathias. *Business process management: concepts, languages, architectures*. Springer Publishing Company, Incorporated, 2010.  
Acceso VPN UGR <http://link.springer.com/book/10.1007%2F978-3-642-28616-2>



- Quién:
  - Analista de proceso.
- Qué:
  - Modelo de proceso.
  - BPMN
- Notación estándar para modelo.
- <http://www.bpmn.org/>

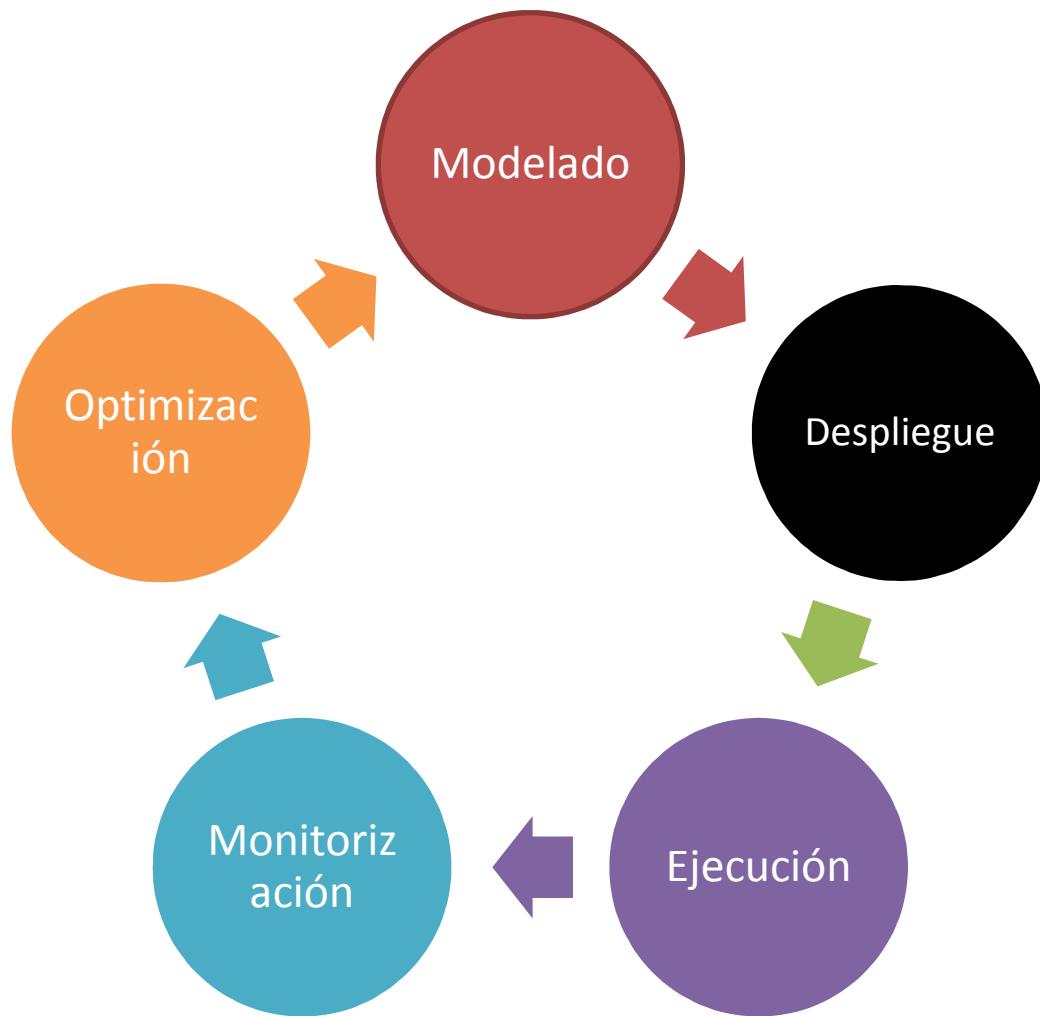




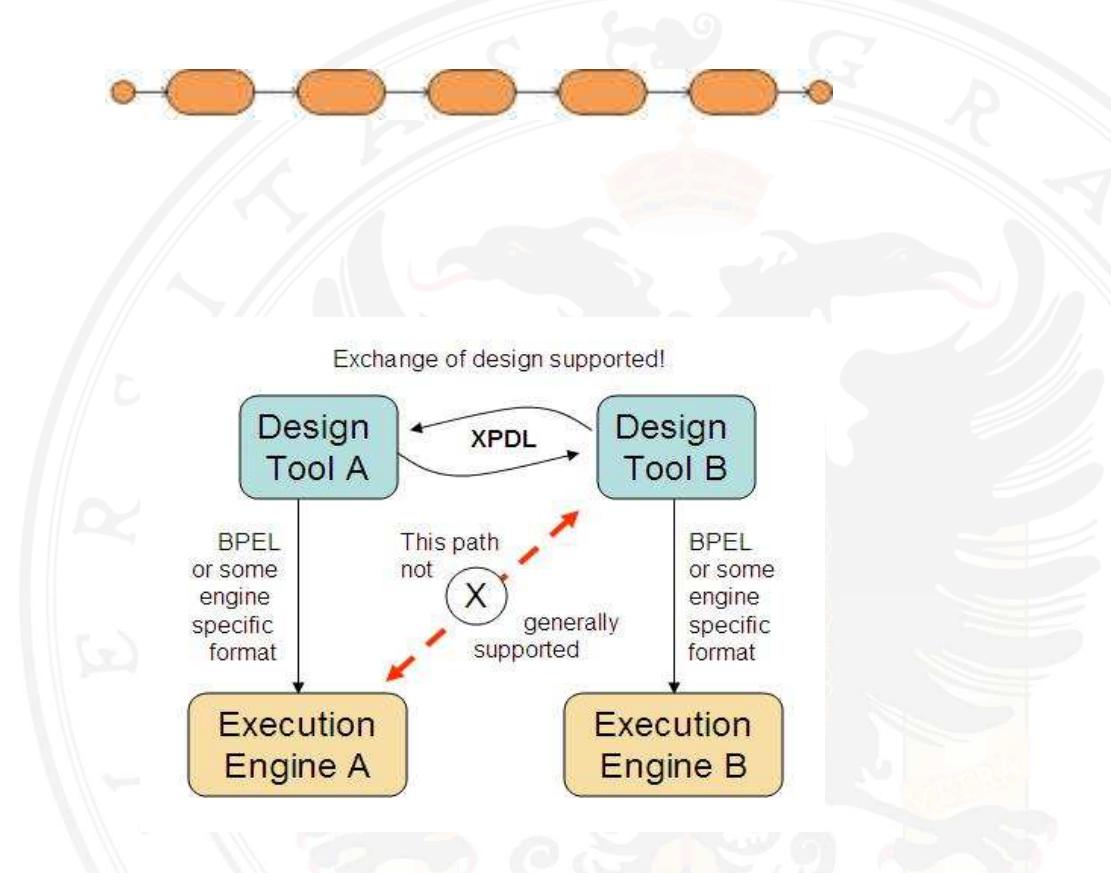
- Quién:
  - Desarrollador.
- Qué:
  - Instancia de proceso
- XPDL: estándar serialización de procesos (instancias o modelos)

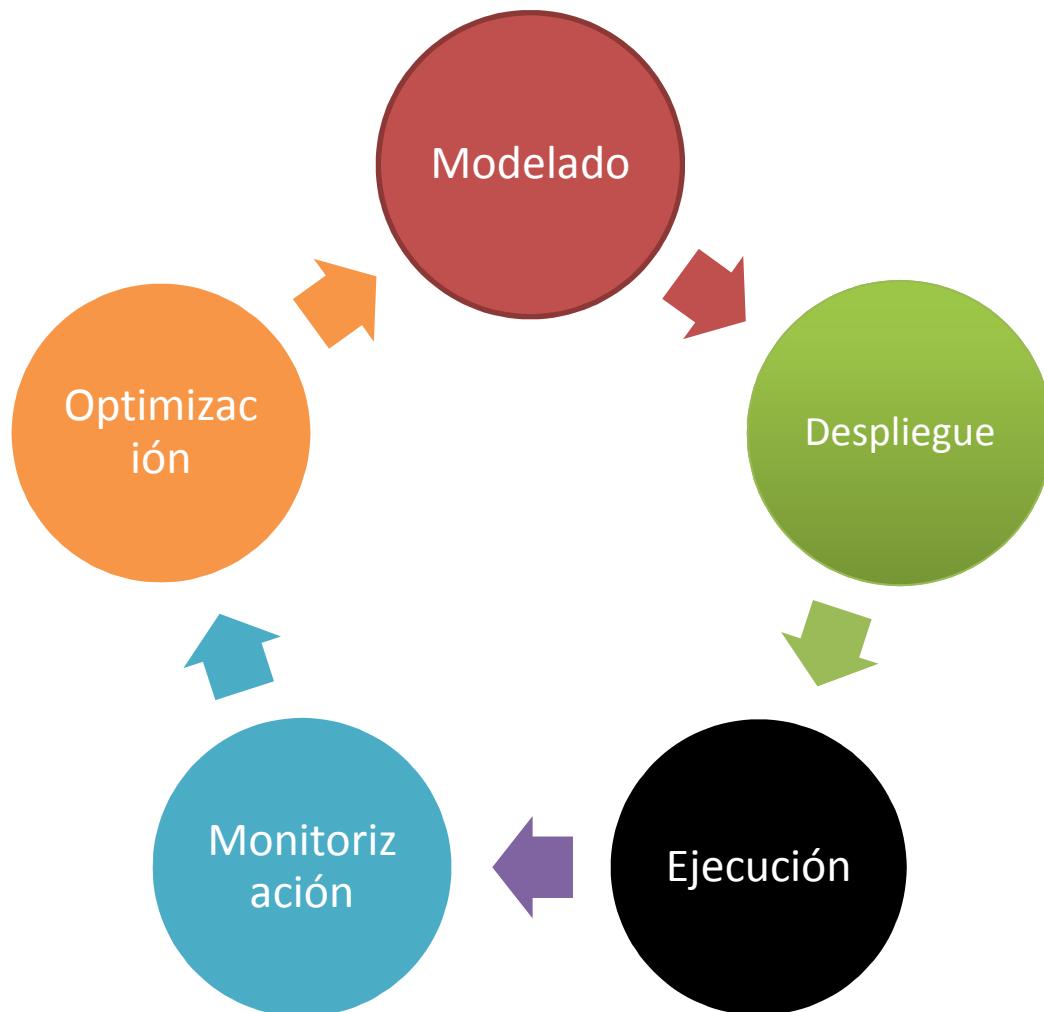


```
<Process>
  <Activities>
    <Activity>...</Activity>
    <Activity>...</Activity>
    <Activity>...</Activity>
    <Activity>...</Activity>
    <Activity>...</Activity>
    <Activity>...</Activity>
  </Activities>
  <Transitions>
    <Transition>...</Transition>
    <Transition>...</Transition>
    <Transition>...</Transition>
    <Transition>...</Transition>
    <Transition>...</Transition>
    <Transition>...</Transition>
  </Transitions>
</Process>
```

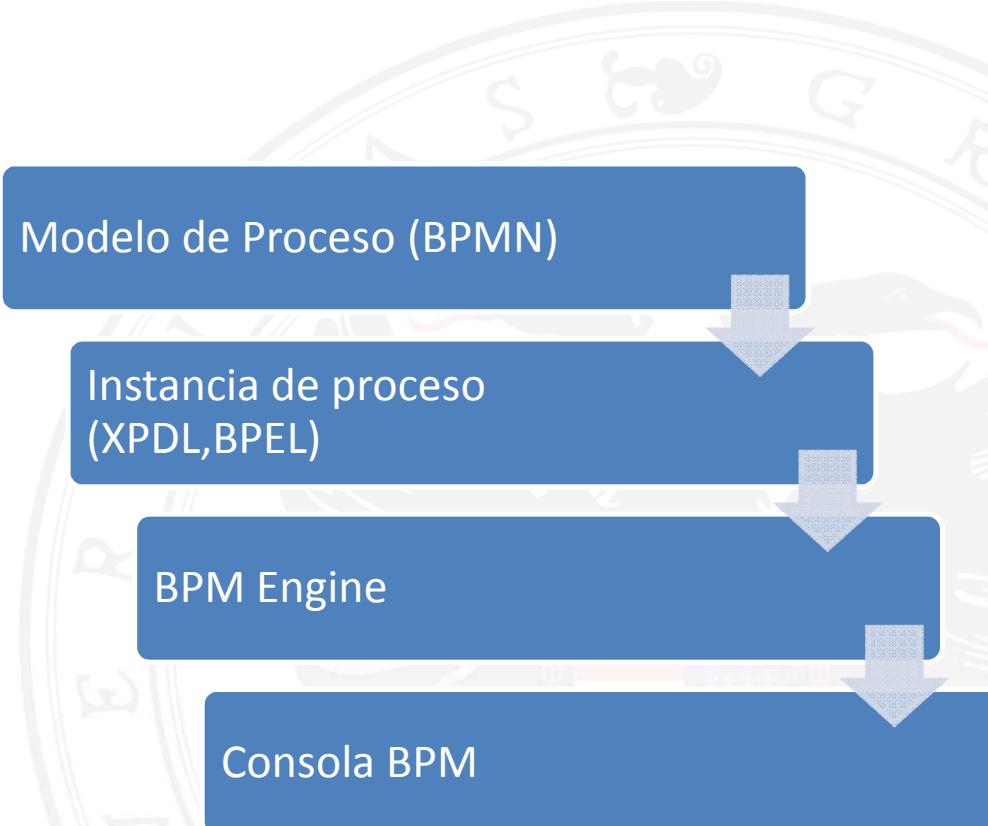


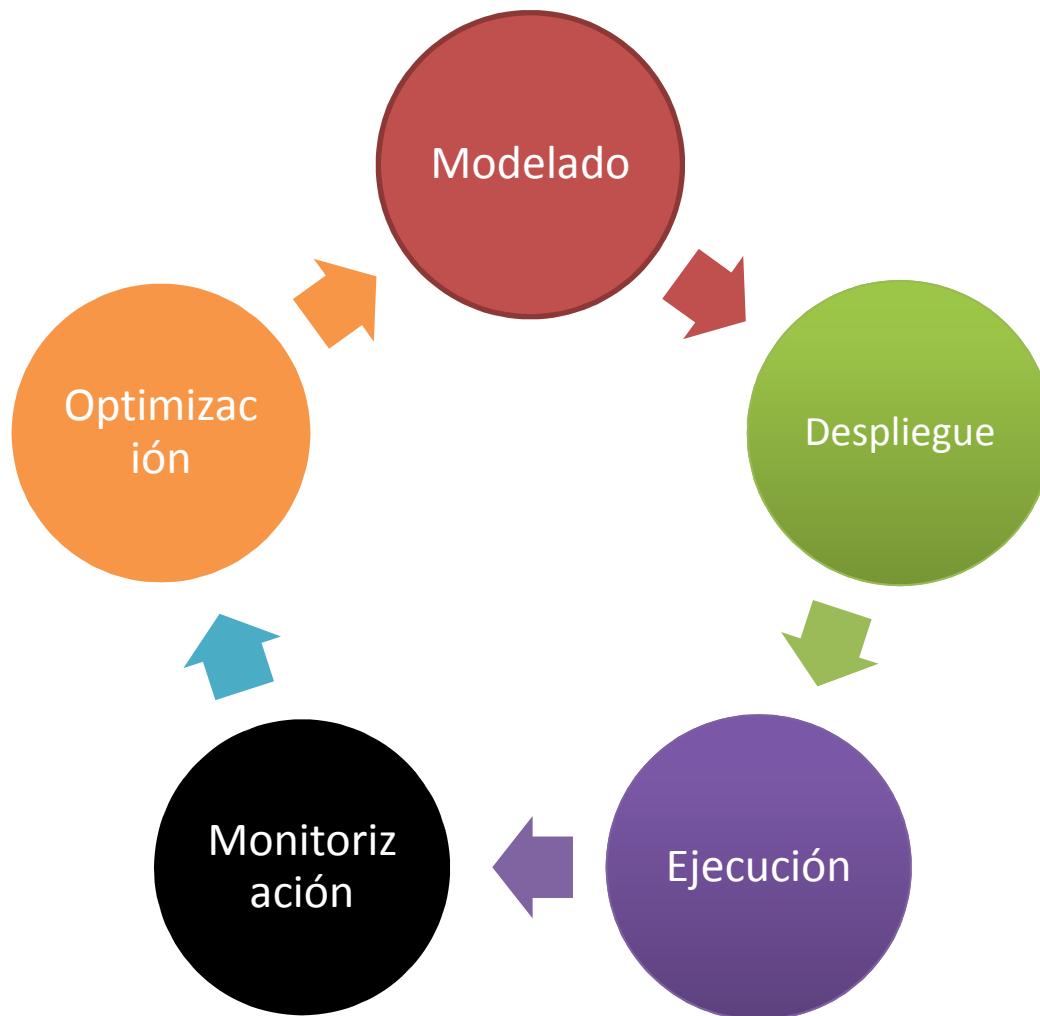
- Quién:
  - Desarrollador.
  - Instancia de proceso
- XPDL: estándar serialización de procesos (instancias o modelos)



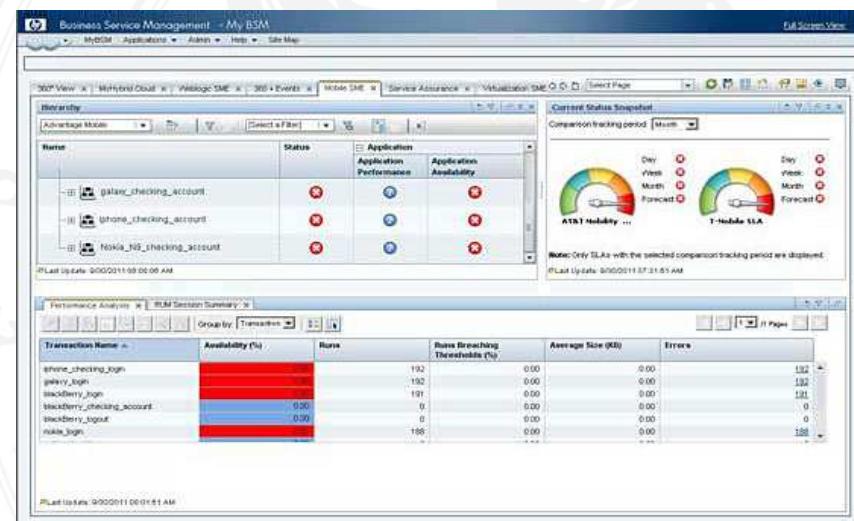


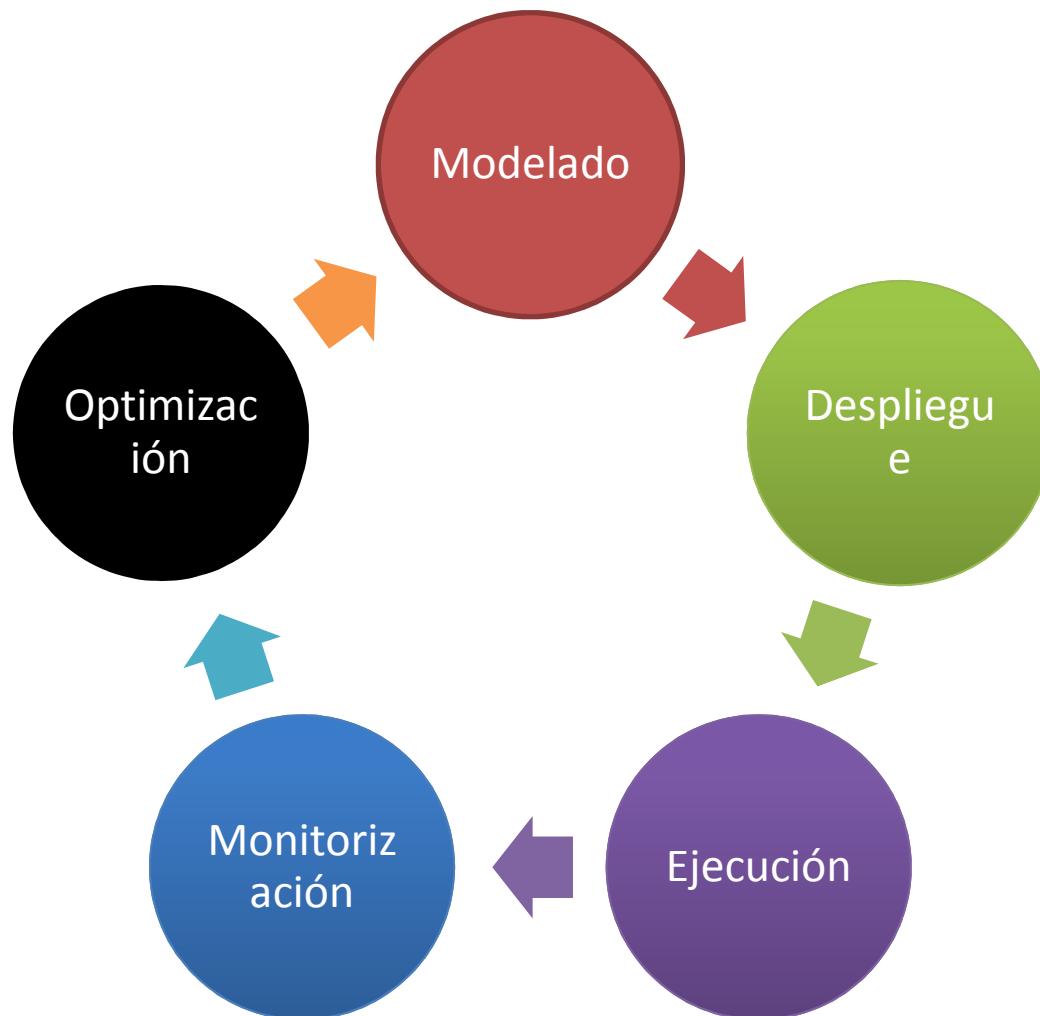
- Quién:
  - BPM engine.
  - Usuarios
- Qué:
  - Ejecución interactiva



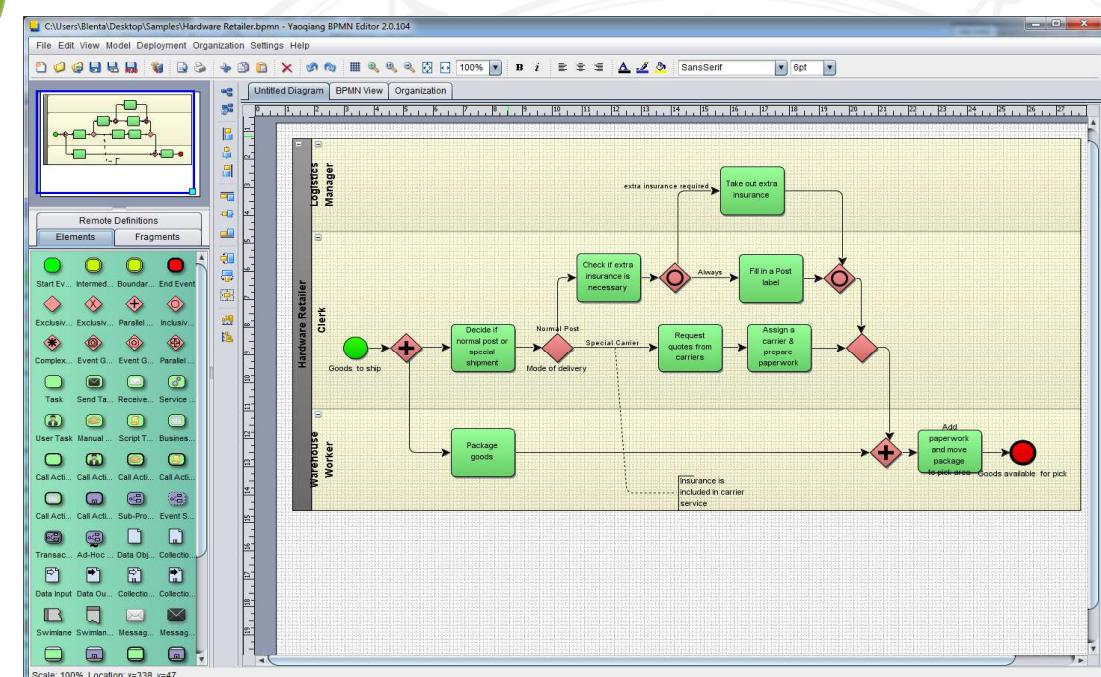


- Quién:
  - BPM monitor
  - Usuarios
- Qué:
  - Estado ejecución, Gestión de excepciones
  - BAM: Business Activity Monitoring:
    - Logs ejecución.
  - KPI: key process indicators
    - Medidas de rendimiento

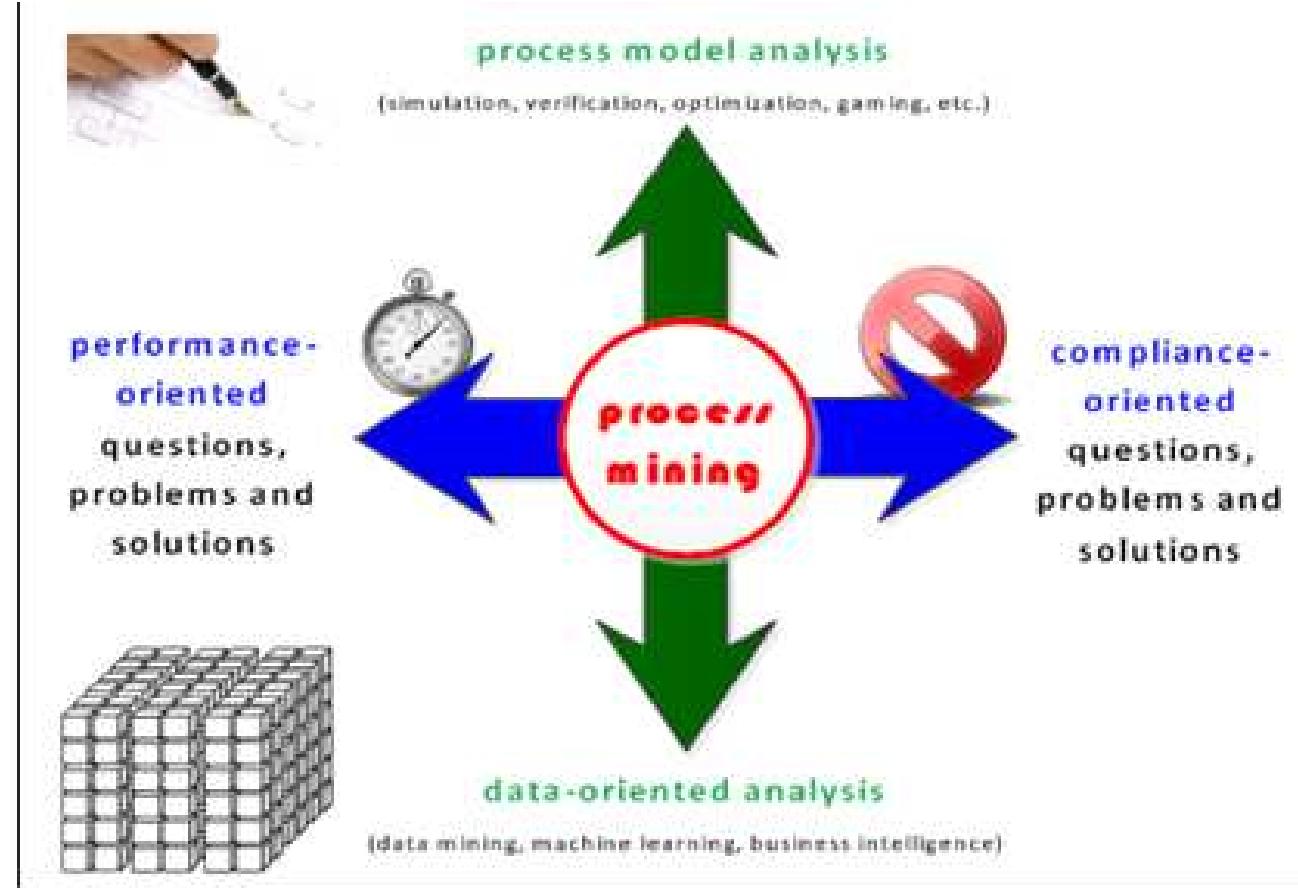




- Quién:
  - Analistas
  - Process Mining/Data Mining
- Qué:
  - Modelo de proceso mejorado

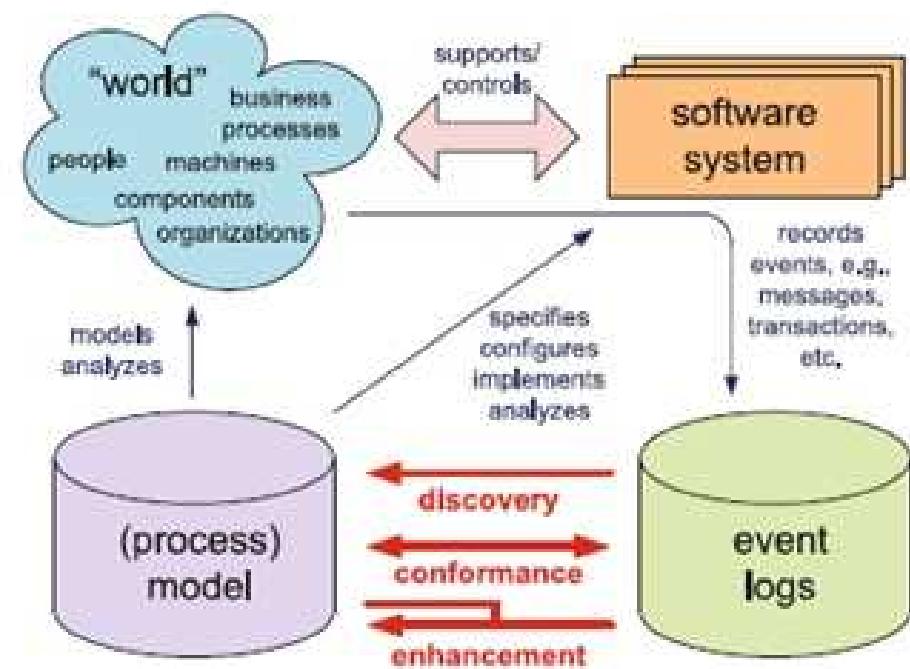


- **Process discovery**
- **Conformance Checking**
- **Enhancement**



- PM ocupa el hueco entre análisis orientado a datos
  - DM, ML, BI
- Y análisis clásico de procesos
  - Simulación, verificación, optimización
- Ofreciendo respuestas sobre rendimiento y conformidad.

- Partiendo de la realidad de lo que ocurre en nuestro mundo.
- Partiendo de un sistema software que usamos para dar soporte o controlar ese entorno
- Registraremos en logs los eventos que se generan con la interacción y procesamiento del sistema.
  - Quejas de clientes, tratamientos de enfermos, ...
- Podremos
  - **Descubrir** modelos de proceso que respondan a conjuntos de eventos.
  - **Comprobar** si una traza de eventos se corresponde con el modelo.
  - **Mejorar** el modelo a partir del análisis de desviaciones o del rendimiento.



- Un registro de una secuencia de eventos que ocurren a lo largo del tiempo.
- Cada evento se refiere a un caso, una actividad y un punto en el tiempo.
- Un log de eventos puede verse como una colección de casos.
- Un caso puede verse como una traza o secuencia de eventos.

## Starting point for process mining:

### Event data

every row is an event  
(here: an exam attempt)

student name	course name	exam date	mark
Peter Jones	Business Information systems	16-1-2014	8
Sandy Scott	Business Information systems	16-1-2014	5
Bridget White	Business Information systems	16-1-2014	9
John Anderson	Business Information systems	16-1-2014	8
Sandy Scott	BPM Systems	17-1-2014	7
Bridget White	BPM Systems	17-1-2014	8
Sandy Scott	Process Mining	20-1-2014	5
Bridget White	Process Mining	20-1-2014	9
John Anderson	Process Mining	20-1-2014	8

case id

activity name

timestamp

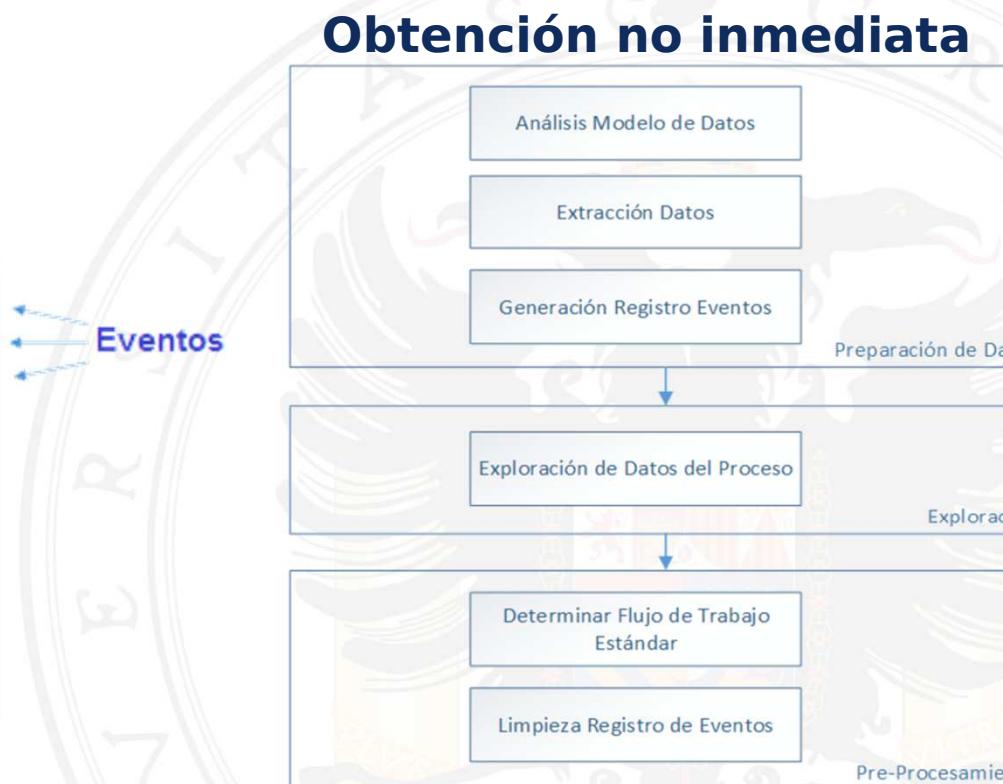
other data

IU/e

- Una base de datos (datos de pacientes en un hospital).
- Un fichero csv o una hoja de cálculo (así será en prácticas).
- Un log de transacciones (cualquier sistema bancario)
- Un ERP o suite BPM (SAP, Oracle,...)
- Un log de mensajes (Linux,...)
- Cualquier API suministrando datos de webs, medios sociales,....

**Instancias**

A	B	C	D	E
1	Identificador de Caso	Actividad	Fecha / Hora	Canal
2	caso_1200	Solicitar	20.02.2002 11:45	cash
3	caso_1200	Registrar	21.02.2002 10:46	sist_trans
4	caso_1202	Solicitar	03.03.2002 15:47	sist_trans
5	caso_1202	Registrar	05.03.2002 09:48	sist_trans
6	caso_1202	Verificar	10.03.2002 11:49	sist_trans
7	caso_1203	Solicitar	10.03.2002 00:50	email
8	caso_1203	Registrar	11.03.2002 10:51	sist_trans
9	caso_1203	Verificar	13.03.2002 11:52	sist_trans
10	caso_1203	Rechazar	16.04.2002 17:53	email
11	caso_1210	Solicitar	15.02.2002 9:54	sist_trans
12	caso_1210	Registrar	15.02.2002 11:55	sist_trans
13	caso_1210	Verificar	20.02.2002 15:56	sist_trans
14	caso_1210	Aprobación	25.02.2002 16:57	sist_trans
15	caso_1210	Liquidacion	01.03.2002 11:58	sist_trans



# Starting point for process mining:

## Event data

every row is an event  
(here: an exam attempt)

student name	course name	exam date	mark
Peter Jones	Business Information systems	16-1-2014	8
Sandy Scott	Business Information systems	16-1-2014	5
Bridget White	Business Information systems	16-1-2014	9
John Anderson	Business Information systems	16-1-2014	8
Sandy Scott	BPM Systems	17-1-2014	7
Bridget White	BPM Systems	17-1-2014	8
Sandy Scott	Process Mining	20-1-2014	5
Bridget White	Process Mining	20-1-2014	9
John Anderson	Process Mining	20-1-2014	8

case id

activity name

timestamp

other data

# Another event log: order handling

order number	activity	timestamp	user	product	quantity
9901	register order	22-1-2014@09.15	Sara Jones	iPhone5S	1
9902	register order	22-1-2014@09.18	Sara Jones	iPhone5S	2
9903	register order	22-1-2014@09.27	Sara Jones	iPhone4S	1
9901	check stock	22-1-2014@09.49	Pete Scott	iPhone5S	1
9901	ship order	22-1-2014@10.11	Sue Fox	iPhone5S	1
9903	check stock	22-1-2014@10.34	Pete Scott	iPhone4S	1
9901	handle payment	22-1-2014@10.41	Carol Hope	iPhone5S	1
9902	check stock	22-1-2014@10.57	Pete Scott	iPhone5S	2
9902	cancel order	22-1-2014@11.08	Carol Hope	iPhone5S	2

case id

activity name

timestamp

resource

other data

TU/e

# Another event log: patient treatment

patient	activity	timestamp	doctor	age	cost
5781	make X-ray	23-1-2014@10.30	Dr. Jones	45	70.00
5541	blood test	23-1-2014@10.18	Dr. Scott	61	40.00
5833	blood test	23-1-2014@10.27	Dr. Scott	24	40.00
5781	blood test	23-1-2014@10.49	Dr. Scott	45	40.00
5781	CT scan	23-1-2014@11.10	Dr. Fox	45	1200.00
5833	surgery	23-1-2014@12.34	Dr. Scott	24	2300.00
5781	handle payment	23-1-2014@12.41	Carol Hope	45	0.00
5541	radiation therapy	23-1-2014@13.57	Dr. Jones	61	140.00
5541	radiation therapy	23-1-2014@13.08	Dr. Jones	61	140.00

case id

activity name

timestamp

resource

other data

TU/e

- No siempre está tan claro
  - Pueden haber varias interpretaciones para caso, actividad.

## Answer: Several possible mappings!

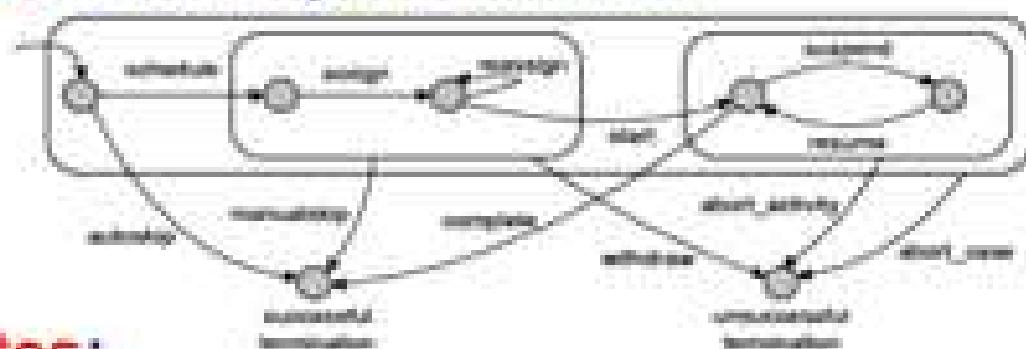
- **Mapping:**
  - a sender ("From"), ——————→ **resource**
  - a set of receivers ("To"), ——————→ **activity name**
  - a subject, ——————→ **case id**
  - a timestamp ("Date"), ——————→ **timestamp**
  - a body, ——————→ **other data**
- **Problems:**
  - Unclear what the cases are (senders, subjects, etc.).
  - Unclear what the activities are.
- **Context and questions needed.**

- No siempre está tan claro:
  - Necesidad de registrar más información:
    - Estados de una actividad: start, complete, suspend,..
    - Atributos de evento propios

## Extensions

- **Transactional information on activity instances:**

An event can represent a **start**, **complete**, **suspend**, **resume**, **abort**, etc.



- **Case versus event attributes:**

- case attributes do not change, e.g., the birth date or gender of a patient,
- event attributes are related to a particular step in the process.

- Un log puede representarse de muchas formas pero hay un estándar para ello.

## XES (eXtensible Event Stream)

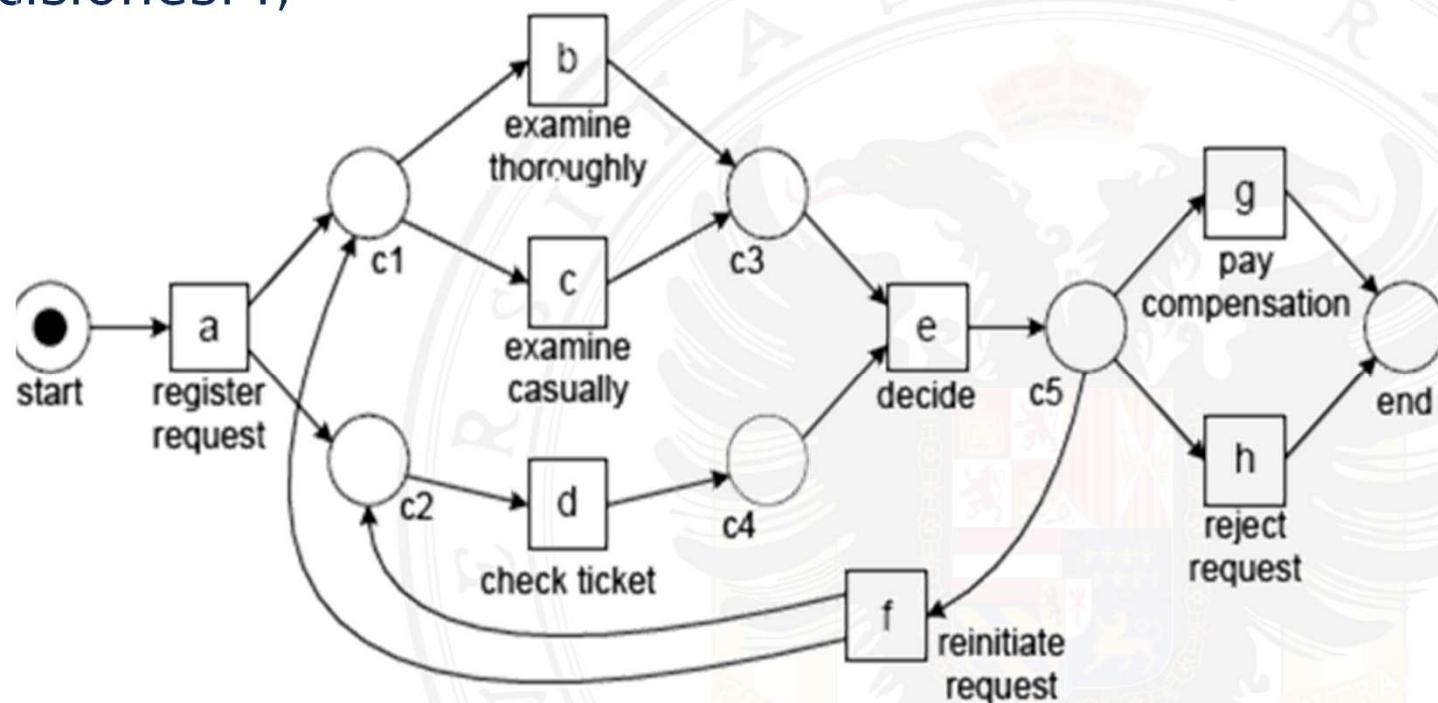
- Adopted by the **IEEE Task Force on Process Mining**.
- The format is supported by tools such as **ProM** and **Disco** (used in this course).
- Predecessors: MXML and SA-MXML.
- Conversion from other formats (**CSV**) is easy if the right data are available.
- **XML syntax** and **OpenXES library** available.
- See [www.xes-standard.org](http://www.xes-standard.org).



- Un modelo de proceso se representa con **un lenguaje o notación de procesos**. Hay un estándar para event logs, pero no hay un estándar para representar modelos de procesos.
- Existen varios lenguajes, con distinta expresividad.
  - petri nets, workflow nets, dependency graphs, causal nets, BPMN, etc.
- Luego los vemos en más detalle, veamos un ejemplo para entender qué relaciones hay entre logs y modelos de proceso.



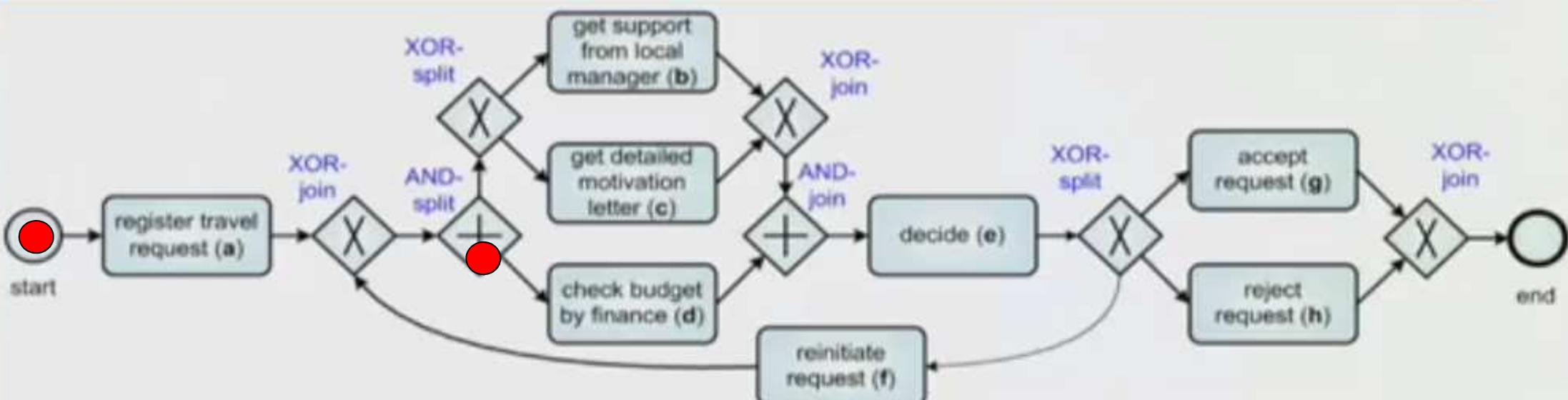
- Caso empieza con a y termina con g o h.
- Actividad d es concurrente con b o c.
- Actividad e tiene que esperar hasta que (d y b) o (d y c) han finalizado.
- Hay tres posibles decisiones: f, g o h.



- Play-out
  - Genero un log a partir de un modelo.
  - Un log puede ser la salida del despliegue de un modelo de proceso en un sistema BPM.
- Play-in
  - Genero un modelo a partir de un log
  - Un log puede ser la entrada de un algoritmo de descubrimiento de procesos.
- Replay
  - Tengo un log y un modelo de proceso y quiero analizar.
  - Un log puede ser la entrada de un modelo de proceso para repetir experimentos off-line.

- A partir del modelo de proceso obtener un log.
  - generar/crear instancias de proceso (secuencias de actividades ejecutables), ejecutarlas y registrar en un log la ejecución de cada actividad como un evento.
  - Se realiza play-out en el típico registro de incidencias o de traza de actividades de cualquier organización (p.ej. Los técnicos de mantenimiento, los operarios de una fábrica, la actividad de uno o varios robots, los alumnos en Prado,...)
  - Esta relación entre logs y modelos de proceso es el uso más extendido de modelos de proceso, es la clave en el ciclo de vida clásico de BPM.
  - Simulación de posibles alternativas de ejecución, a partir de un modelo conocido (análisis what-if).

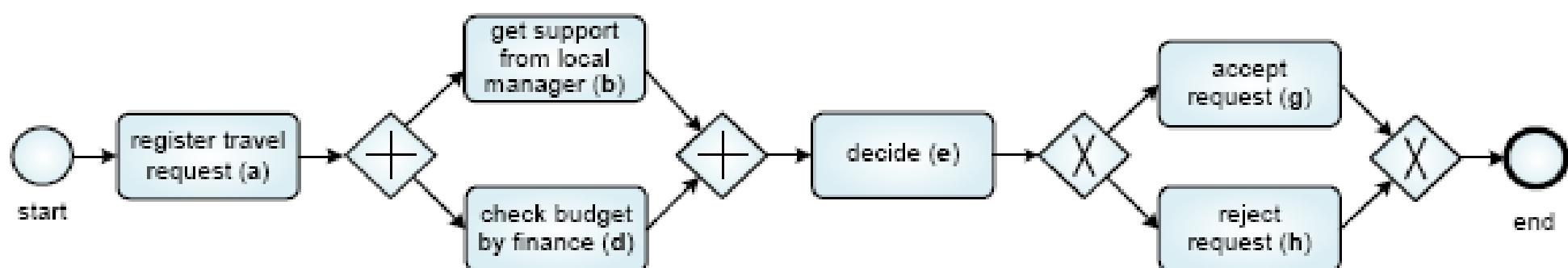
## Play Out: A possible scenario



Case	Activity	Timestamp	Resource
121	Register travel request	14/04@8:00	Pepe
121	Get support from l. mng	14/04@9:00	Pepe
121	Check budget by finance	14/04@9:00	Pepe
121	Decide	14/04@9:00	Pepe
121	Accept request	14/04@9:00	Pepe

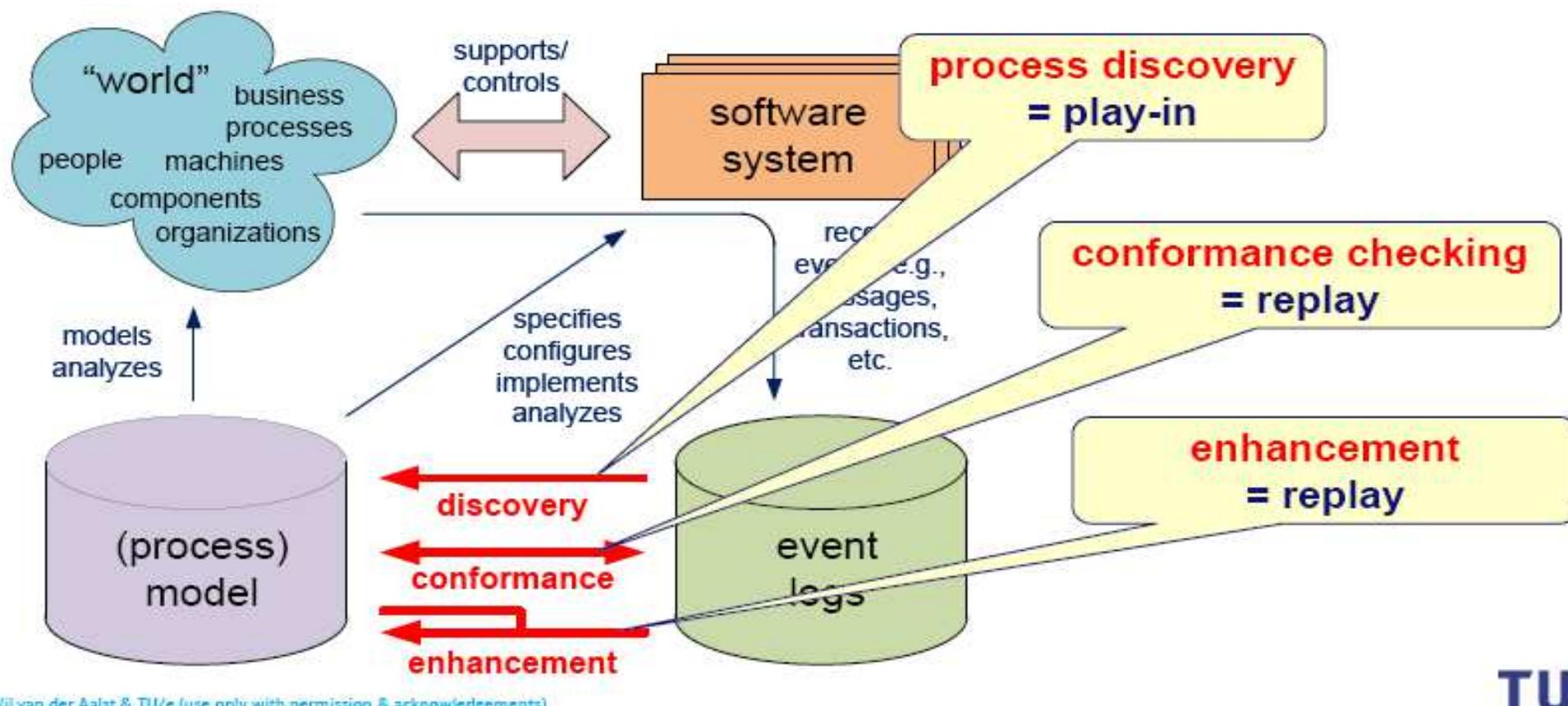
- A partir de un log de eventos obtener un modelo de proceso.
  - Inferir un modelo automáticamente desde trazas de ejecución.
  - Modelar un proceso desde ejemplos.
- ¡¡¡No necesitamos modelar a mano!!!
- En esto nos detendremos fundamentalmente.

abdeg adbeg adbeg adbeh  
abdeh abdeh abdeg abdeh  
abdeh abdeh abdeh adbeh  
adbeh adbeh adbeh adbeh



- Reproducir la realidad (de los logs de eventos) sobre un modelo de proceso ya conocido (descubierto o a mano)
- **Conformance checking.**
  - Ver si hay trazas que se ajustan
  - Ver qué trazas no se ajustan y donde se producen desviaciones.
- **Análisis de rendimiento**
  - Podemos hacer análisis sobre uso de recursos o tiempo.
  - Podemos tomar una traza y registrar el tiempo que tarda cada actividad en ejecutarse
  - También podremos ver los retrasos entre cada actividad
  - Esto lo podemos hacer para varias y obtendríamos frecuencias, caminos más frecuentes,...
- A partir de estos análisis puedo **mejorar mi proceso.**

En resumen:



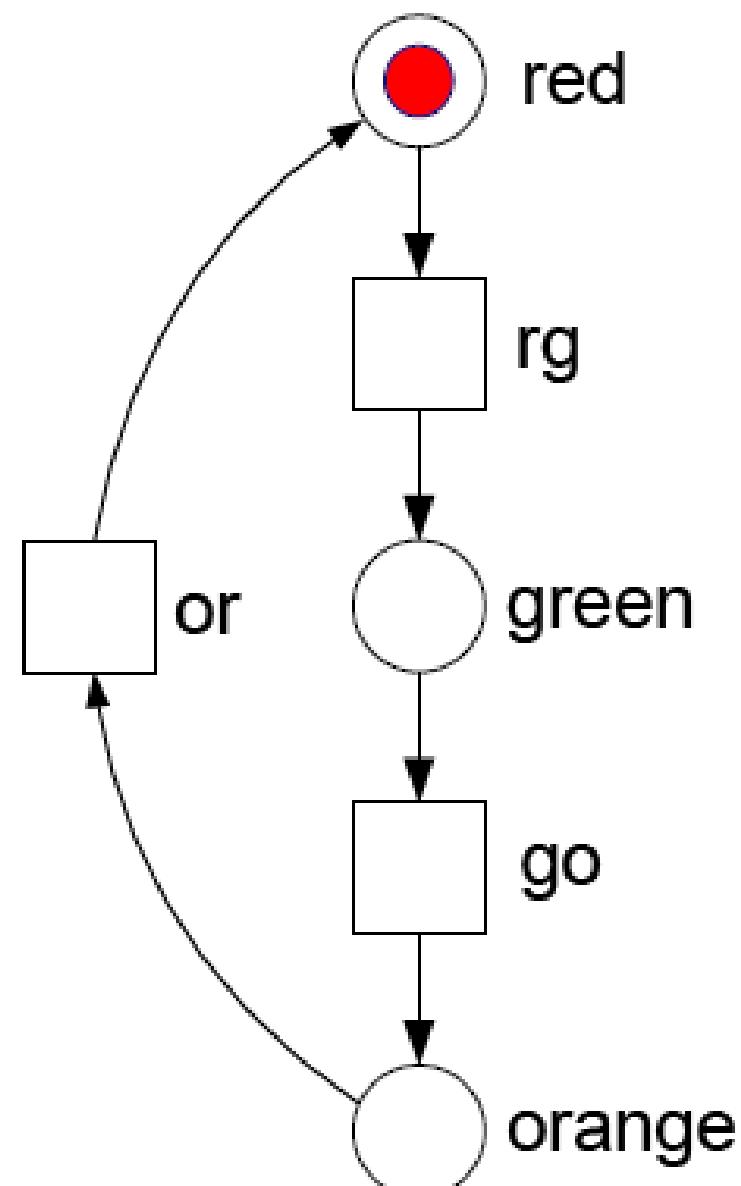
Una vez conocida la representación de logs de eventos y las relaciones que pueden tener con un modelo de proceso, veamos un ejemplo de lenguaje de representación de modelos.

Las **Redes de Petri** son una representación formal de modelos de proceso que se usa como salida del algoritmo alpha, un algoritmo básico de process discovery.

- Un proceso muy simple:
  - Un semáforo puede estar en 3 estados posibles
  - Entre cada par de estados hay una transición.



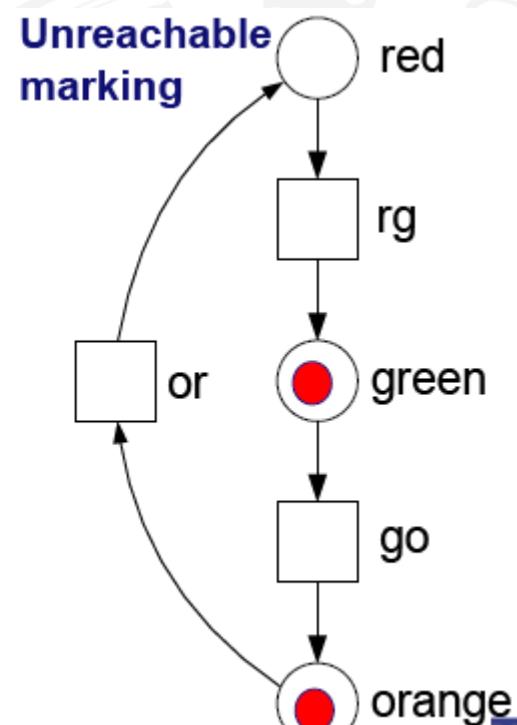
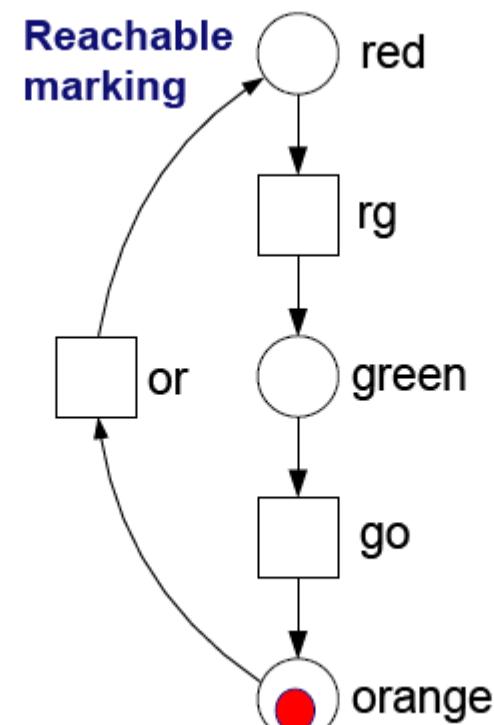
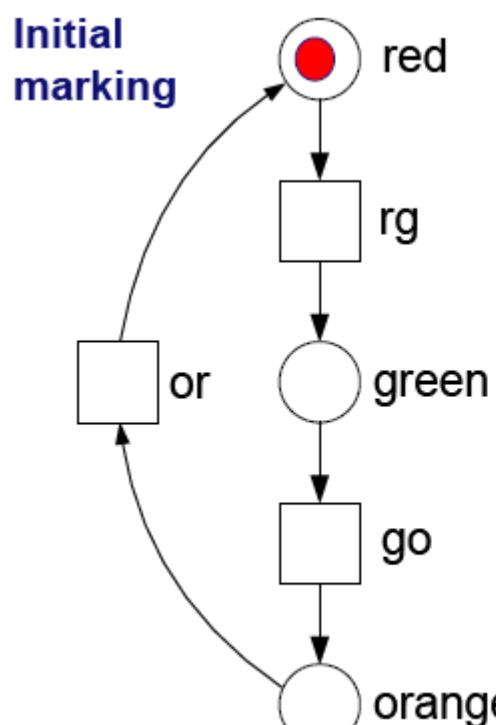
- Una red cuyos nodos pueden ser:
  - Lugares (condiciones activación)
  - Transiciones (actividades)
- Cuyos Arcos
  - conectan lugares y transiciones representando el flujo de control.
- Lugares pueden tener tokens
  - Representando una condición cierta.
- Transiciones
  - Consumen tokens de sus lugares de entrada
  - Producen tokens para sus lugares de salida.



- Veamos algunos conceptos básicos
  - Marcado
  - Activación y disparo
  - Qué puede representarse con una Red de Petri.
  - Ejemplo de dos semáforos
    - No determinista
    - Determinista



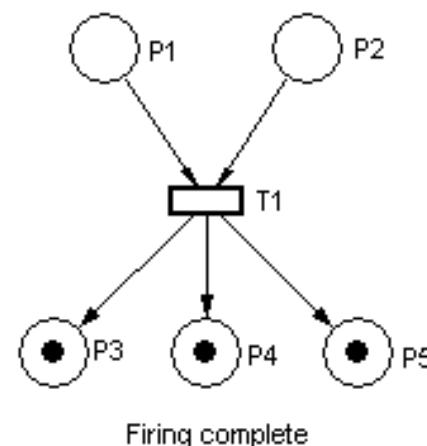
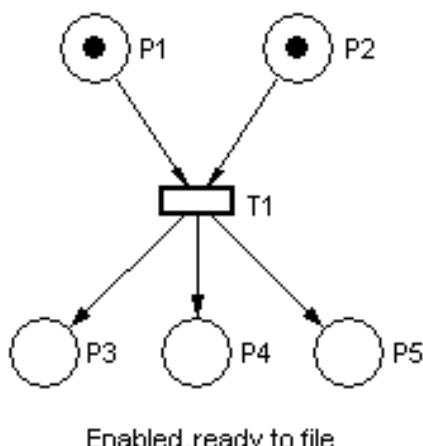
- **Marcado**
  - Un marcado representa una instantánea de la PN, e.d., una configuración o estado en que se pueda encontrar.
- **Tipos de marcados:**
  - Inicial,
  - Alcanzable
  - Inalcanzable



© Wil van der Aalst & TU/e (use only with permission & acknowledgements)

## Activación:

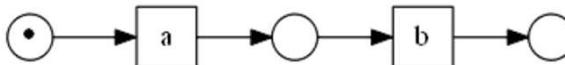
- Una transición está activada si cada lugar de entrada contiene un token.
- Disparo
  - Un transición activada puede dispararse consumiendo **un token** desde cada lugar de entrada y produciendo un token para cada lugar de salida.



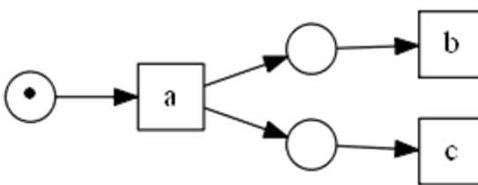
- Para representar formalmente modelos de proceso.
  - Pueden representar los patrones de proceso más frecuentemente usados
  - Ver siguiente transparencia.
- Pueden traducirse a otras notaciones de proceso
  - Dado un modelo en una notación gráfica, puedo analizar propiedades de la bondad de ese modelo
  - Dada una PN obtenida de un algoritmo de aprendizaje, puedo traducirla a otro lenguaje
    - Para rediseñar
    - Para ejecutar directamente.

## COMMON WORKFLOW PATTERNS

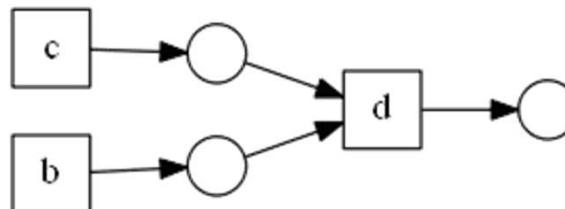
<http://www.workflowpatterns.com/>



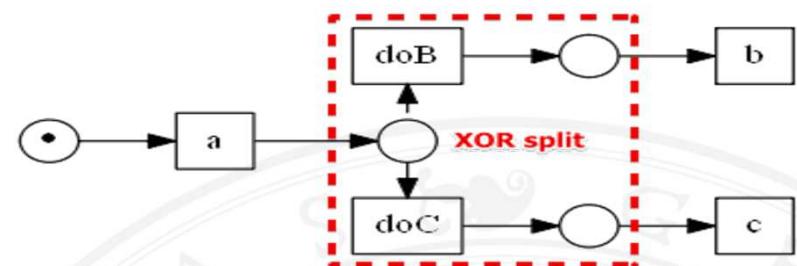
Secuencia: ejecutar a después de b



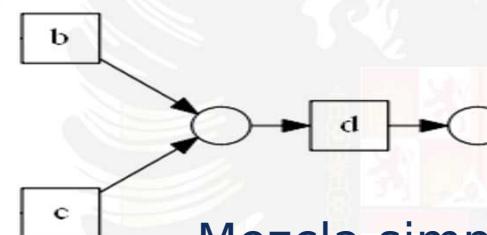
Split paralelo (AND-split): b y c se ejecutan en paralelo después de a



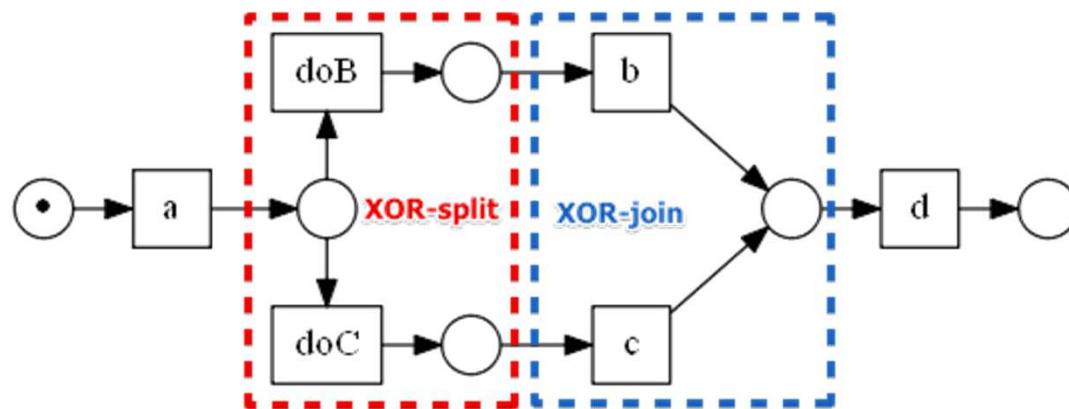
Sincronización (AND-join): ejecutar d después del completado de c y b



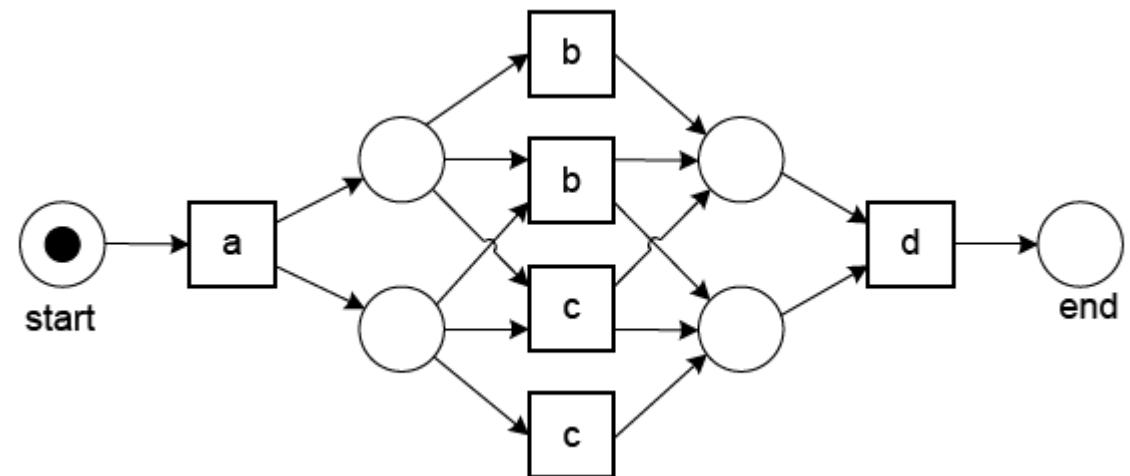
Eleción exclusiva(XOR-split): después de a, ejecutar o bien b o bien c



Mezcla simple(XOR-join): ejecutar d después de acaben b o c.

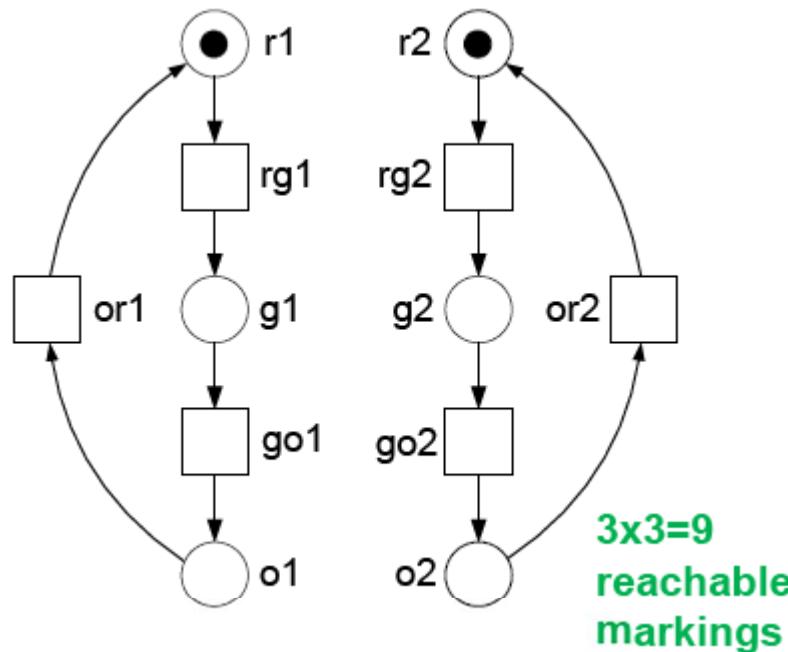


Enrutado alternativo (estructura condicional) combinando XOR-split con XOR-join



OR-split/join: después de a, ejecutar, o bien b, o bien ambas (en paralelo).

- **Play out** es comprobar todos los caminos de los tokens, empezando por start y acabando en end.
- **Play in** es determinar la Red de Petri que genera un conjunto de trazas de entrada.
- **Replay** permite detectar dónde están los problemas en la Red de Petri cuando reciba como entrada una traza real que no puede ser reproducida, e indicaciones sobre cómo repararla.



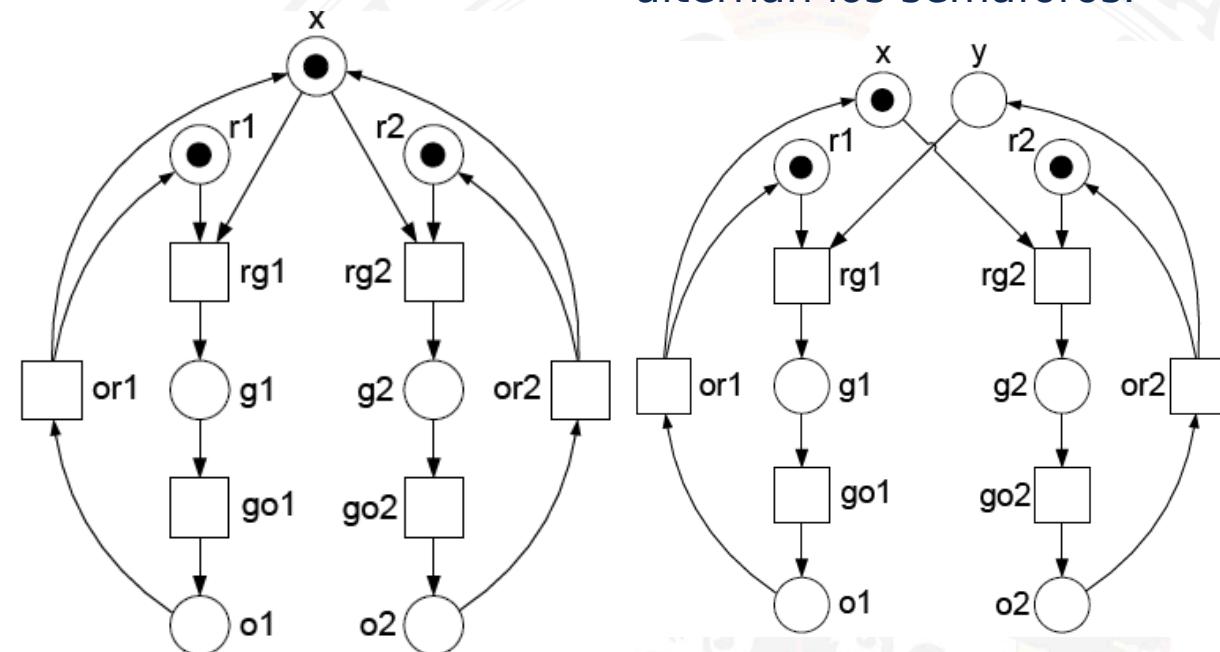
¿Cómo la podemos hacer segura?



Esta Red de Petri no es segura porque puede llevar a situaciones que no queremos que se produzcan

Modelo no determinista

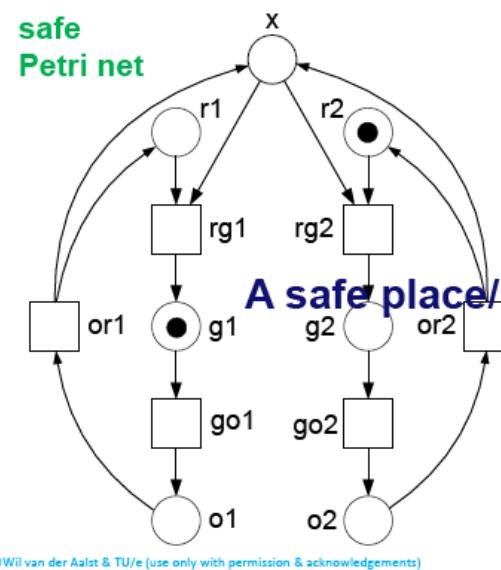
Modelo determinista en el que decidimos cómo se alternan los semáforos.



- Análisis de la calidad/bondad de un modelo
  - Podemos usar las PN para estudiar la calidad de modelos obtenidos
  - ¿Puedo garantizar que en mi modelo
    - ¿Las actividades concurrentes hacen buen uso de recursos compartidos?.
    - ¿No hay actividades que se queden bloqueadas?
    - ¿No hay actividades inútiles (que nunca se ejecutan)?

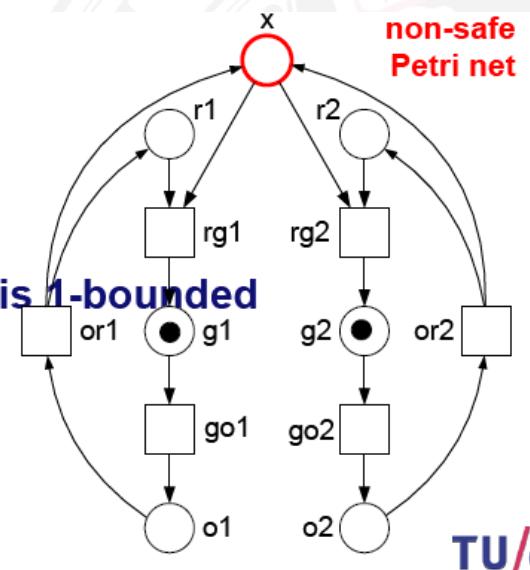
## Análisis de la calidad de mi modelo

- Podemos usar las PN para estudiar la calidad de modelos obtenidos
- ¿Puedo garantizar que en mi modelo
  - las actividades concurrentes hacen buen uso de recursos compartidos?
    - **Seguridad:** una red de petri es segura si garantiza que no se puede alcanzar un marcado en el que los lugares tengan más de un token.
  - No hay actividades que se queden bloqueadas?
  - No hay actividades inútiles (que nunca se ejecutan)?



A safe place/Petri net is 1-bounded

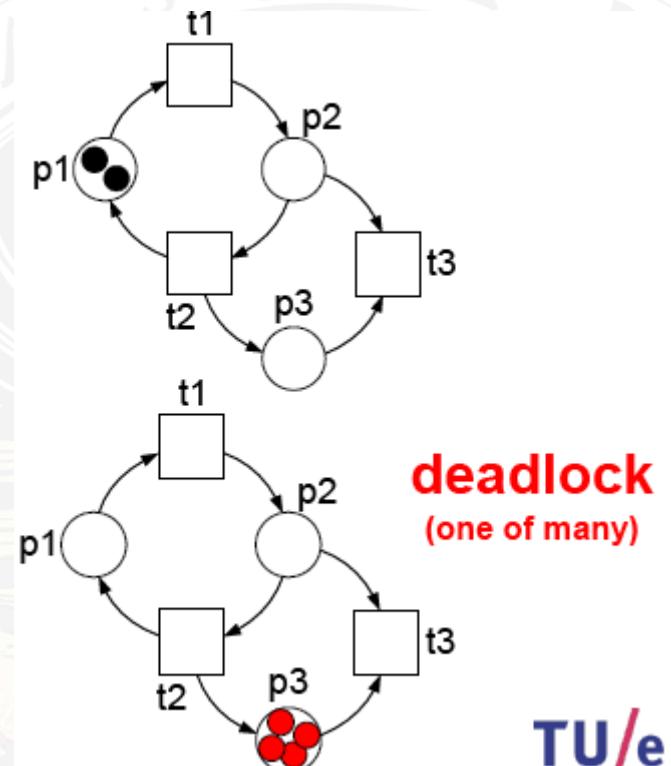
© Wil van der Aalst & TU/e (use only with permission & acknowledgements)



TU/e

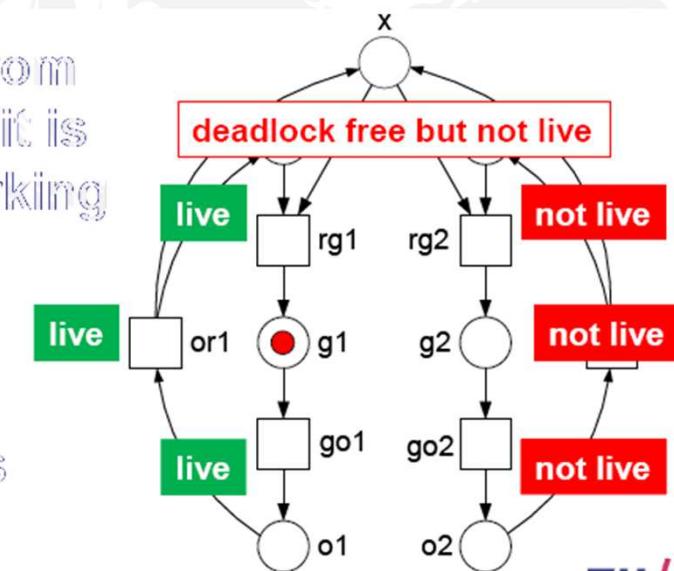
## Análisis de la calidad de mi modelo

- Podemos usar las PN para estudiar la calidad de modelos obtenidos
- ¿Puedo garantizar que en mi modelo
  - las actividades concurrentes hacen buen uso de recursos compartidos?
    - **Seguridad**: una red de petri es segura si garantiza que no se puede alcanzar un marcado en el que los lugares tengan más de un token.
  - No hay actividades que se queden bloqueadas?
    - **Libre de bloqueos**.: una PN está libre de bloqueos si cada marcado alcanzable activa al menos una transición.
  - No hay actividades inútiles (que nunca se ejecutan)?

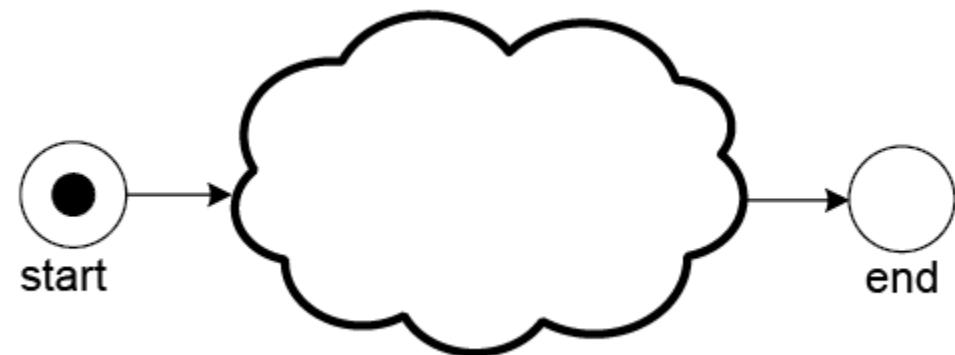


## Análisis de la calidad de mi modelo

- Podemos usar las PN para estudiar la calidad de modelos obtenidos
  - ¿Puedo garantizar que en mi modelo
    - las actividades concurrentes hacen buen uso de recursos compartidos?
      - **Seguridad:** una red de petri es segura si garantiza que no se puede alcanzar un marcado en el que los lugares tengan más de un token.
    - No hay actividades que se queden bloqueadas?
      - **Libre de bloqueos.:** una PN está libre de bloqueos si cada marcado alcanzable activa al menos una transición.
    - No hay actividades inútiles (que nunca se ejecutan)?
      - **Vivacidad:** una PN es viva si todas sus transiciones son vivas, e.d., si garantiza que siempre es posible alcanzar un marcado que activa cada transición.



- Las WF-nets son un caso particular de PN y el formalismo básico usado en BPM para hacer análisis basado en modelos.
- Una WF-net tiene un solo lugar de inicio y un solo lugar de fin y todos los otros nodos están en un camino desde inicio hasta fin (representa un proceso bien formado)

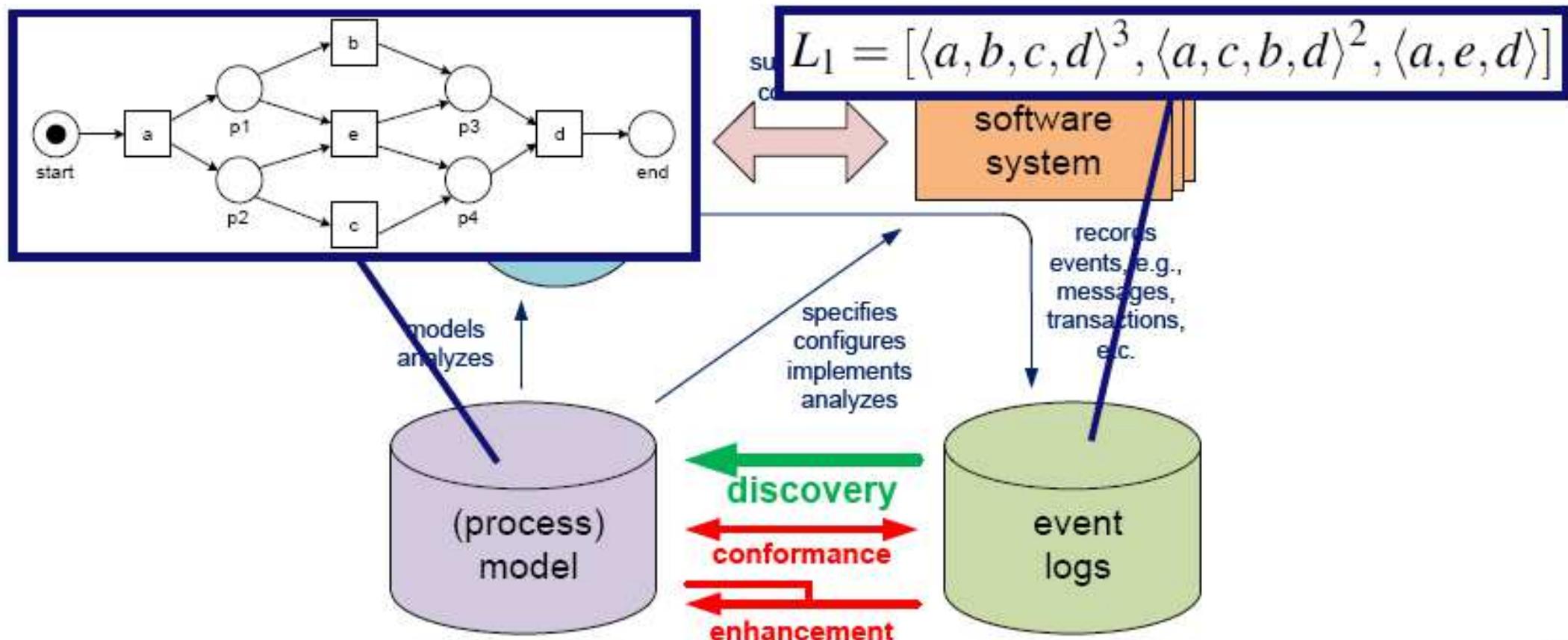


- Una WF-net es correcta si se cumplen las siguientes propiedades.
  - Seguridad: los lugares no pueden tener múltiples tokens a la vez.
  - Completado propio: Si el lugar final está marcado, todos los lugares están vacíos (end-to-end process).
  - Opción de completado: Siempre es posible alcanzar un marcado que marca sólo el lugar final.
  - Ausencia de partes muertas: Para cualquier transición hay una secuencia de disparo que la activa.

- Resumiendo, las Redes de Petri (WF-nets) son un lenguaje para representar modelos de procesos.
  - Lugares, transiciones, arcos, tokens, marcado y activación.
- Una vez representado un modelo de proceso como una Red de Petri puedo llevar a cabo análisis basado en modelos:
  - Verificar y validar formalmente, a partir de las propiedades de seguridad, vivacidad y libre de interbloqueos.
  - Análisis de rendimiento:
    - *Simular (Play-out)*
    - *Replay (comprobar cómo se adapta el modelo a una traza real)*.
- Limitaciones del análisis basado en modelos:
  - Verificación y análisis se basan en que se dispone de modelos de muy buena calidad.
  - Cuando un modelo de proceso y la realidad tienen poco en común, el análisis basado en modelos pierde sentido.
  - A menudo hay un alineamiento pobre entre modelos hechos a mando y realidad.
  - Process Mining trata de afrontar estos problemas estableciendo una conexión directa entre modelos y los datos de eventos actuales.



**Descubrimiento de procesos:** hacer corresponder un **log de eventos** con un modelo de proceso tal que sea representativo del **comportamiento observado**.



## Process discovery is like learning a language: By example



## Process discovery is like learning a language: By example



## Process discovery is like learning a language: By example



## Process discovery is like learning a language: By example

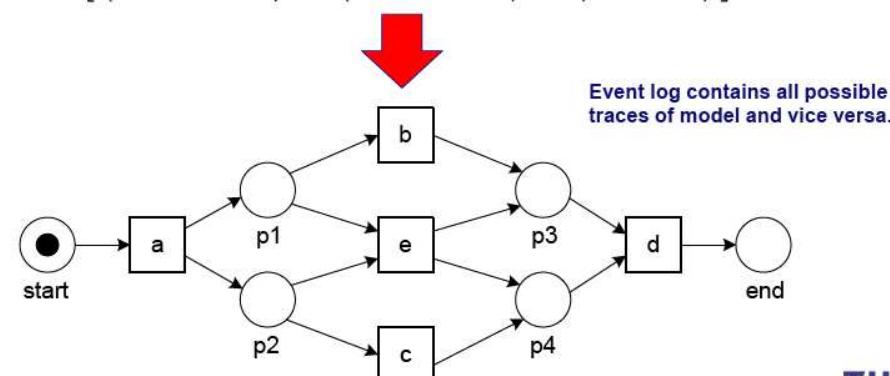


sentence  $\cong$  trace in event log  
language  $\cong$  process model ...

## Algoritmo $\alpha$

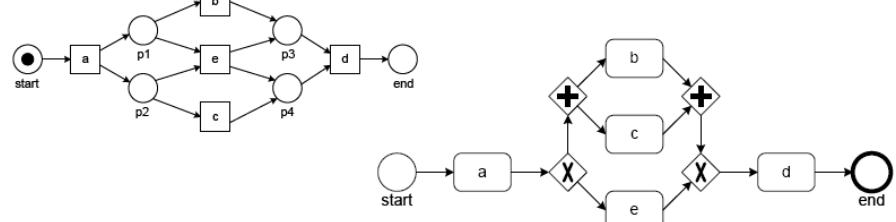
- Entrada: Un log simplificado
- Salida: Un modelo de proceso
- Objetivo: generar un modelo (una WF-net) de proceso que pueda explicar secuencias de eventos del log.

$$L_1 = [\langle a, b, c, d \rangle^3, \langle a, c, b, d \rangle^2, \langle a, e, d \rangle]$$



- La notación usada como salida no es muy relevante.
- El algoritmo obtiene una WF-net que puede luego traducirse a alguna notación más usada en BPM
- Lo importante son los tipos de relaciones entre actividades que pueden detectarse y por tanto cómo de expresivo es el modelo obtenido.

Notation is less relevant (e.g. BPMN)



$$L_1 = [\langle a, b, c, d \rangle^3, \langle a, c, b, d \rangle^2, \langle a, e, d \rangle]$$

© Wil van der Aalst & TU/e (use only with permission & acknowledgements)

- Conocer qué forma tiene un log simplificado
- Conocer las 4 relaciones que puede haber entre los eventos del log
- Conocer cómo representar el “footprint” de un log como una matriz.
- Conocer qué patrones pueden ser descubiertos



- Simplificación: sólo control de flujo, sólo consideramos actividades.

order number	activity	timestamp	user	product	quantity
9901	register order	22-1-2014@09.15	Sara Jones	iPhone5S	1
9902	register order	22-1-2014@09.18	Sara Jones	iPhone5S	2
9903	register order	22-1-2014@09.27	Sara Jones	iPhone4S	1
9901	check stock	22-1-2014@09.49	Pete Scott	iPhone5S	1
9901	ship order	22-1-2014@10.11	Sue Fox	iPhone5S	1
9903	check stock	22-1-2014@10.34	Pete Scott	iPhone4S	1
9901	handle payment	22-1-2014@10.41	Carol Hope	iPhone5S	1
9902	check stock	22-1-2014@10.57	Pete Scott	iPhone5S	2

[<register\_order, check\_stock, ship\_order, handle\_payment>,  
 <register\_order, check\_stock, cancel\_order>,  
 <register\_order, check\_stock>, ...]

- Log de eventos simple

$$L_1 = [\langle a, b, c, d \rangle^3, \langle a, c, b, d \rangle^2, \langle a, e, d \rangle]$$

- Log de eventos: un multiconjunto de trazas (la misma traza puede aparecer varias veces)
- Una traza es una secuencia ordenada de nombres de actividades, anotada con la frecuencia de aparición en el log.

- **Sucesión directa:  $x > y$**   
sií para algún caso x está seguida directamente por y.

A partir del log de eventos simple podemos extraer distintas relaciones fundamentales entre las actividades del log.

$$L_1 = [\langle a, b, c, d \rangle^3, \langle a, c, b, d \rangle^2, \langle a, e, d \rangle]$$

a>b
a>c
a>e
b>c
b>d
c>b
c>d
e>d

- Sucesión directa:  $x > y$  sií para algún caso  $x$  está seguida directamente por  $y$ .
- Causalidad:  $x \rightarrow y$   
sií  $x > y$  and  $\text{not}(y > x)$

A partir del log de eventos simple podemos extraer distintas relaciones fundamentales entre las actividades del log.

$$L_1 = [\langle a, b, c, d \rangle^3, \langle a, c, b, d \rangle^2, \langle a, e, d \rangle]$$

a>b  
a>c  
a>e  
b>c  
b>d  
c>b  
c>d  
e>d

$a \rightarrow b$   
 $a \rightarrow c$   
 $a \rightarrow e$   
 $b \rightarrow d$   
 $c \rightarrow d$   
 $e \rightarrow d$

- Sucesión directa:  $x > y$  sií para algún caso  $x$  está seguida directamente por  $y$ .
- Causalidad:  $x \rightarrow y$  sií  $x > y$  and  $\text{not}(y > x)$
- Paralelismo:  $x \parallel y$  sií  $x > y$  and  $y < x$

A partir del log de eventos simple podemos extraer distintas relaciones fundamentales entre las actividades del log.

$$L_1 = [\langle a, b, c, d \rangle^3, \langle a, c, b, d \rangle^2, \langle a, e, d \rangle]$$

$a > b$   
 $a > c$   
 $a > e$   
 $b > c$   
 $b > d$   
 $c > b$   
 $c > d$   
 $e > d$

$a \rightarrow b$   
 $a \rightarrow c$   
 $a \rightarrow e$   
 $b \rightarrow d$   
 $c \rightarrow d$   
 $e \rightarrow d$

$b \parallel c$   
 $c \parallel b$

- Sucesión directa:  $x > y$  sií para algún caso  $x$  está seguida directamente por  $y$ .
- Causalidad:  $x \rightarrow y$  sií  $x > y$  and  $\text{not}(y > x)$
- Paralelismo:  $x \parallel y$  sií  $x > y$  and  $y < x$
- Elección (choice):  $x \# y$  sií  $\text{not}(x > y)$  and  $\text{not}(y > x)$

A partir del log de eventos simple podemos extraer distintas relaciones fundamentales entre las actividades.

$$L_1 = [\langle a, b, c, d \rangle^3, \langle a, c, b, d \rangle^2, \langle a, e, d \rangle]$$

$a > b$   
 $a > c$   
 $a > e$   
 $b > c$   
 $b > d$   
 $c > b$   
 $c > d$   
 $e > d$

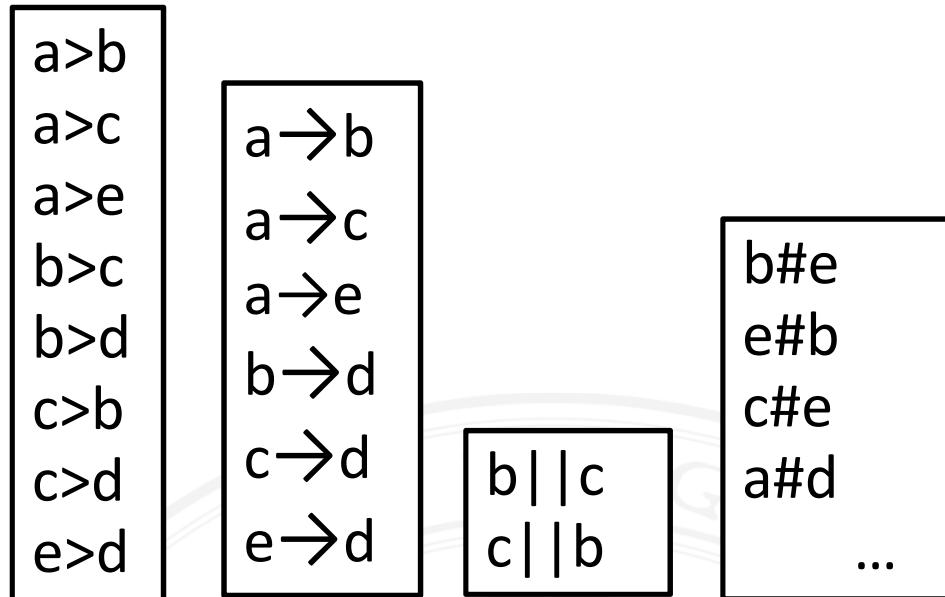
$a \rightarrow b$   
 $a \rightarrow c$   
 $a \rightarrow e$   
 $b \rightarrow d$   
 $c \rightarrow d$   
 $e \rightarrow d$

$b \parallel c$   
 $c \parallel b$

$b \# e$   
 $e \# b$   
 $c \# e$   
 $a \# d$   
...

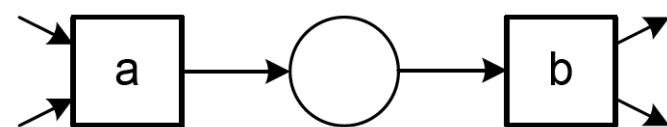
**Footprint:** Una representación de las relaciones entre actividades como una matriz de adyacencia anotada.

$$L_1 = [\langle a, b, c, d \rangle^3, \langle a, c, b, d \rangle^2, \langle a, e, d \rangle]$$

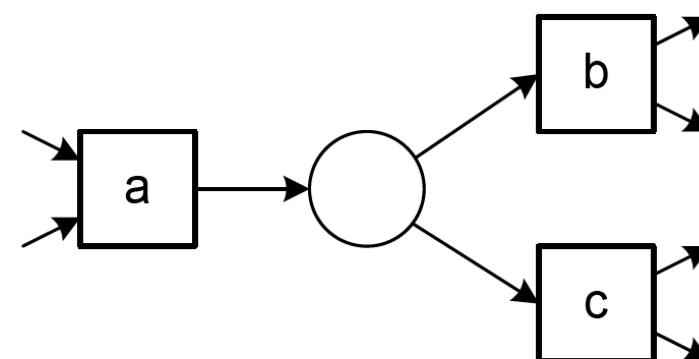


	<i>a</i>	<i>b</i>	<i>c</i>	<i>d</i>	<i>e</i>
<i>a</i>	# $L_1$	$\rightarrow_{L_1}$	$\rightarrow_{L_1}$	# $L_1$	$\rightarrow_{L_1}$
<i>b</i>	$\leftarrow_{L_1}$	# $L_1$	$L_1$	$\rightarrow_{L_1}$	# $L_1$
<i>c</i>	$\leftarrow_{L_1}$	$L_1$	# $L_1$	$\rightarrow_{L_1}$	# $L_1$
<i>d</i>	# $L_1$	$\leftarrow_{L_1}$	$\leftarrow_{L_1}$	# $L_1$	$\leftarrow_{L_1}$
<i>e</i>	$\leftarrow_{L_1}$	# $L_1$	# $L_1$	$\rightarrow_{L_1}$	# $L_1$

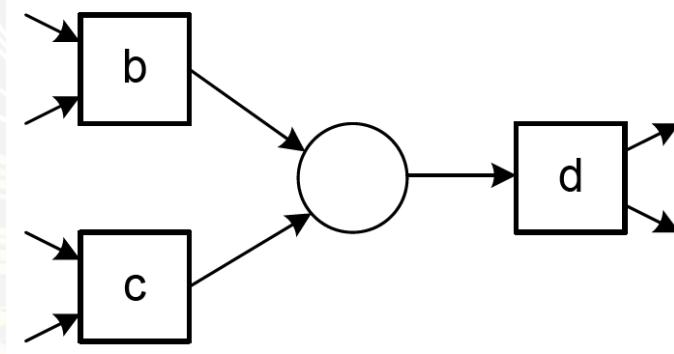
*Las relaciones fundamentales permiten expresar patrones de proceso comunes que aparecen en cualquier representación de modelos de proceso.*



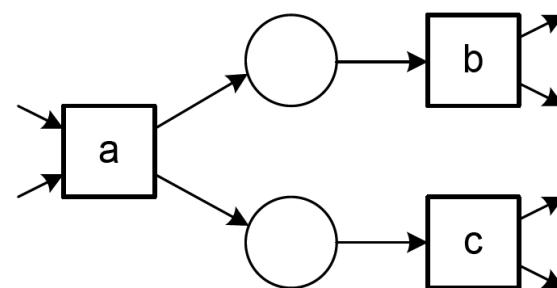
(a) sequence pattern:  $a \rightarrow b$



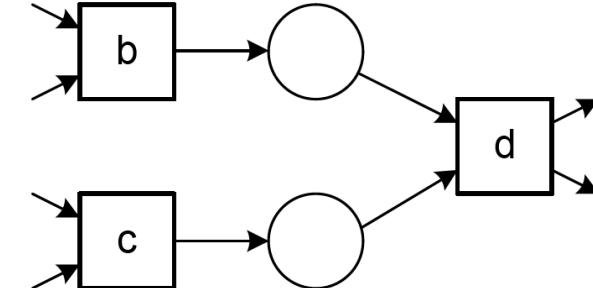
(b) XOR-split pattern:  
 $a \rightarrow b$ ,  $a \rightarrow c$ , and  $b \# c$



(c) XOR-join pattern:  
 $b \rightarrow d$ ,  $c \rightarrow d$ , and  $b \# c$



(d) AND-split pattern:  
 $a \rightarrow b$ ,  $a \rightarrow c$ , and  $b \parallel c$



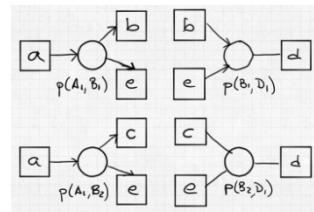
(e) AND-join pattern:  
 $b \rightarrow d$ ,  $c \rightarrow d$ , and  $b \parallel c$

[Consultar](#)

<http://www.workflowpatterns.com/>

case id	event id	properties			
		timestamp	activity	resource	exit
1	35654429	30-12-29 0:11:00	register request	Pete	50
	35654424	30-12-29 0:10:00	examine thoroughly	Sara	400
	35654425	05-01-29 1:05:38	check ticket	Mike	100
	35654426	06-01-29 1:31:38	decide	Sara	200
2	35654427	07-01-29 1:54:24	reject request	Pete	200
	35654483	30-12-29 0:11:32	register request	Mike	50
	35654485	30-12-29 0:12:12	check ticket	Mike	100
	35654487	30-12-29 0:14:16	examine carefully	Pete	400
	35654488	05-01-29 1:11:22	decide	Sara	200
3	35654489	08-01-29 1:12:08	pay compensation	Ellen	200
	35654521	30-12-29 0:14:32	register request	Pete	50
	35654522	30-12-29 0:15:06	examine carefully	Mike	400
	35654524	30-12-29 0:16:34	check ticket	Ellen	100
	35654525	06-01-29 1:09:18	decide	Sara	200
4	35654526	06-01-29 1:12:18	reject request	Sara	200

	<i>a</i>	<i>b</i>	<i>c</i>	<i>d</i>	<i>e</i>
<i>a</i>	# <i>L</i> <sub>1</sub>	→ <i>L</i> <sub>1</sub>	→ <i>L</i> <sub>1</sub>	# <i>L</i> <sub>1</sub>	→ <i>L</i> <sub>1</sub>
<i>b</i>	← <i>L</i> <sub>1</sub>	# <i>L</i> <sub>1</sub>	<i>L</i> <sub>1</sub>	→ <i>L</i> <sub>1</sub>	# <i>L</i> <sub>1</sub>
<i>c</i>	← <i>L</i> <sub>1</sub>	<i>L</i> <sub>1</sub>	# <i>L</i> <sub>1</sub>	→ <i>L</i> <sub>1</sub>	# <i>L</i> <sub>1</sub>
<i>d</i>	# <i>L</i> <sub>1</sub>	← <i>L</i> <sub>1</sub>	← <i>L</i> <sub>1</sub>	# <i>L</i> <sub>1</sub>	← <i>L</i> <sub>1</sub>
<i>e</i>	← <i>L</i> <sub>1</sub>	# <i>L</i> <sub>1</sub>	# <i>L</i> <sub>1</sub>	→ <i>L</i> <sub>1</sub>	# <i>L</i> <sub>1</sub>



Procesar el log para simplificarlo

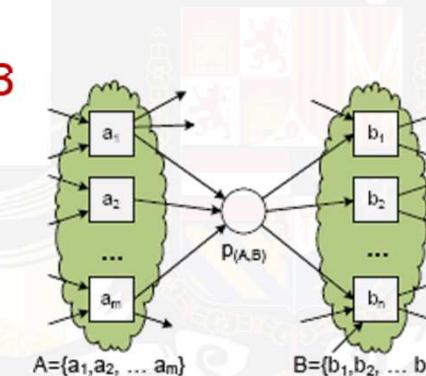
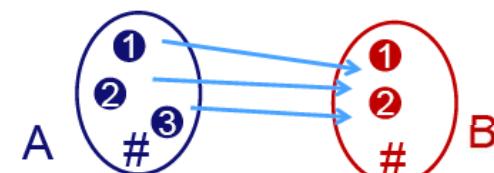
Crear el footprint

Determinar conjuntos de transiciones que estén relacionados

Determinar los lugares a partir de estos conjuntos

Conectar lugares y transiciones.

case id	trace
1	$\langle a, b, d, e, h \rangle$
2	$\langle a, d, c, e, g \rangle$
3	$\langle a, c, d, e, f, b, d, e, g \rangle$
4	$\langle a, d, b, e, h \rangle$
5	$\langle a, c, d, e, f, d, c, e, f, c, d, e, h \rangle$
6	$\langle a, c, d, e, g \rangle$
...	...



1. Determinar el conjunto de actividades diferentes del log.
  1.  $T = \{a, b, c, d, e\}$
2. Determinar los conjuntos de primeras actividades en cada traza y de últimas actividades de cada traza
  1.  $I = \{a\}$
  2.  $F = \{d\}$
3. Crear el footprint

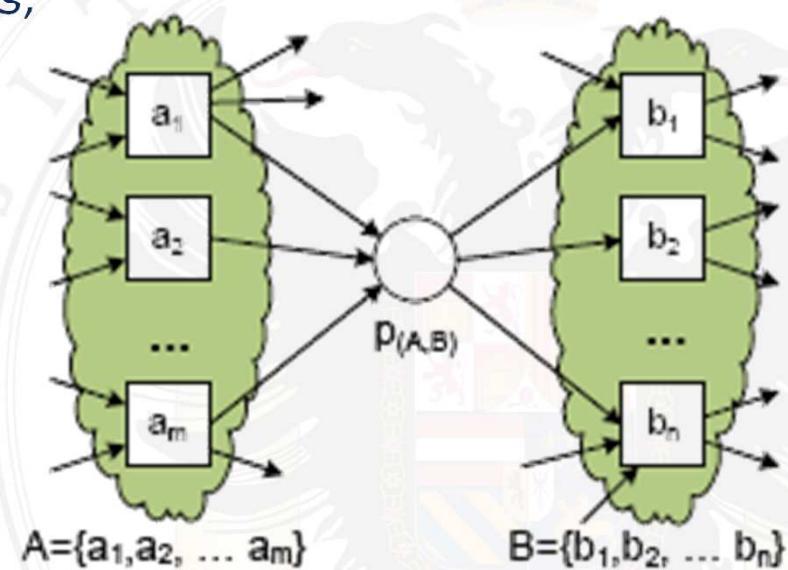
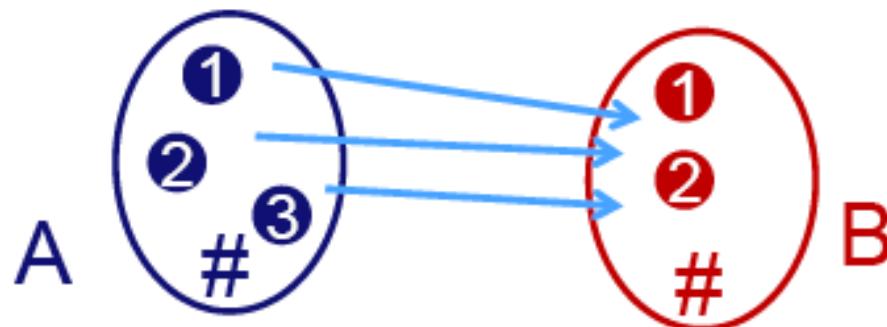
	$a$	$b$	$c$	$d$	$e$
$a$	$\#_{L_1}$	$\rightarrow_{L_1}$	$\rightarrow_{L_1}$	$\#_{L_1}$	$\rightarrow_{L_1}$
$b$	$\leftarrow_{L_1}$	$\#_{L_1}$	$\parallel_{L_1}$	$\rightarrow_{L_1}$	$\#_{L_1}$
$c$	$\leftarrow_{L_1}$	$\parallel_{L_1}$	$\#_{L_1}$	$\rightarrow_{L_1}$	$\#_{L_1}$
$d$	$\#_{L_1}$	$\leftarrow_{L_1}$	$\leftarrow_{L_1}$	$\#_{L_1}$	$\leftarrow_{L_1}$
$e$	$\leftarrow_{L_1}$	$\#_{L_1}$	$\#_{L_1}$	$\rightarrow_{L_1}$	$\#_{L_1}$

los siguientes pasos se basan en determinar los lugares de la WF-net.

4. Etapa clave para descubrir los lugares de la WF-net: encontrar un conjunto de **PARES DE CONJUNTOS** de actividades (**A, B**), tales que, PARA CADA PAR, cada elemento de A está causalmente relacionado con **TODOS LOS** elementos de B, y los elementos de A son independientes entre sí (ninguno de A es seguido por otro de A, **incluso por sí mismo**), y los de B también.

- $X_{L_1} = \{\{\{a\}, \{b\}\}, \{\{a\}, \{c\}\}, \{\{a\}, \{e\}\}, \{\{a\}, \{b,e\}\}, \{\{a\}, \{c,e\}\}, \{\{b\}, \{d\}\}, \{\{c\}, \{d\}\}, \{\{e\}, \{d\}\}, \{\{b,e\}, \{d\}\}, \{\{c,e\}, \{d\}\}, \}$

(*ojo, pueden haber pares no maximales, e.d, pares que cumplan las propiedades, y que estén incluidos en otros cjos.*)

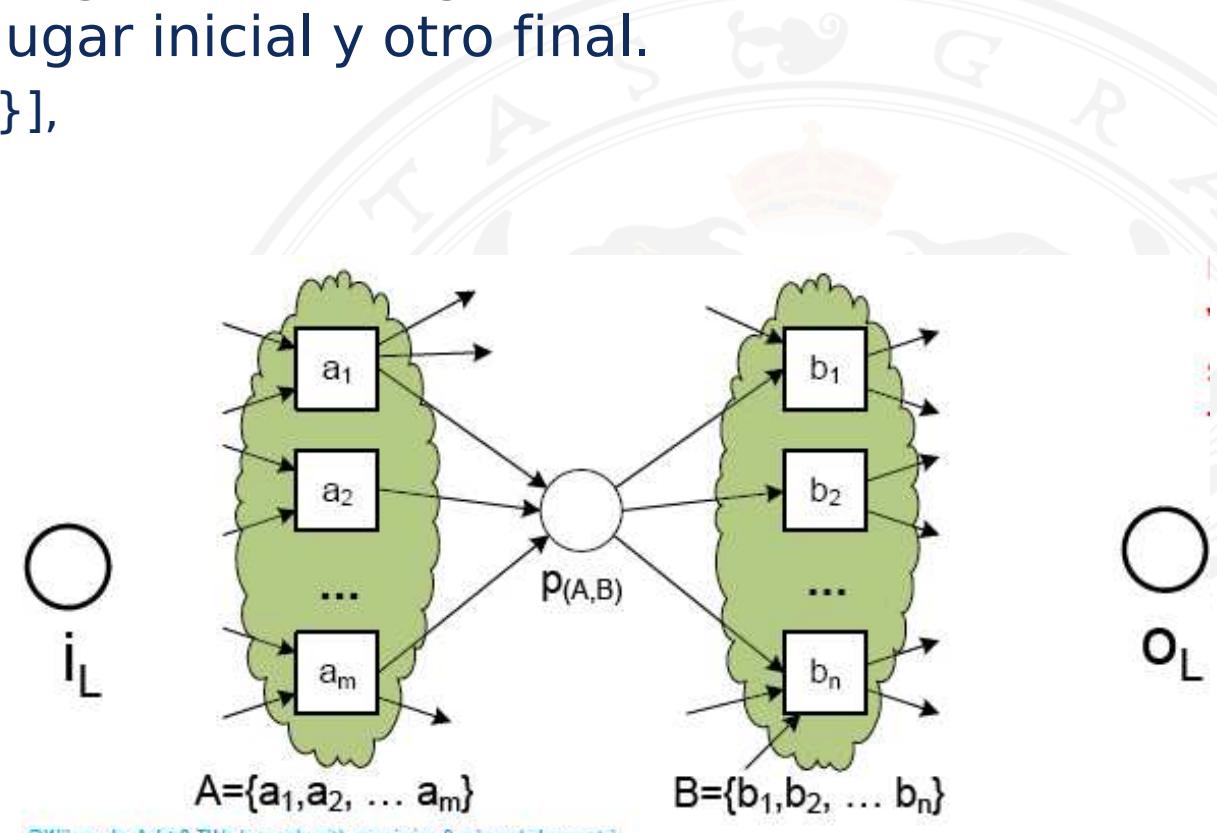


5. Encontrar  $Y_L$  mediante **eliminación** de los pares de  $X_L$  no maximales,e.d, aquellos pares  $(A', B')$  que se pueden obtener a partir de otro  $(A, B)$  y que mantienen las propiedades.

1.  $Y_{L1} = \{\underline{\{[a], [b]\}}, \underline{\{[a], [c]\}}, \underline{\{[a], [e]\}}, \{[a], \{b, e\}\}, \{[a], \{c, e\}\}, \underline{\{[b], [d]\}}, \underline{\{[c], [d]\}}, \underline{\{[e], [d]\}}, \{[b, e], \{d\}\}, \{[c, e], \{d\}\}, \}$

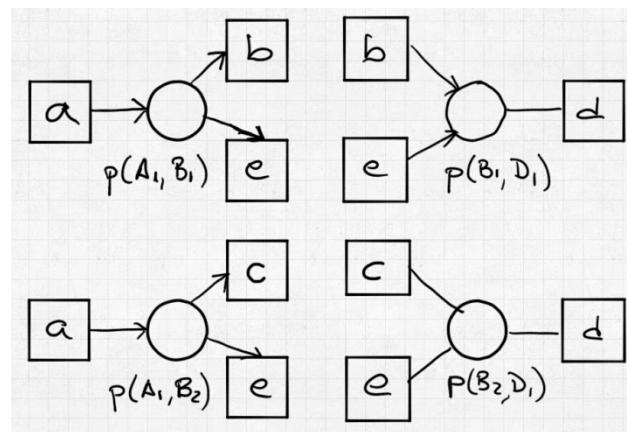
6. Determinar el conjunto  $P_L$  de lugares de la WF-net : cada elemento  $(A, B)$  de  $Y_L$  es un lugar. Para asegurar la estructura de flujo de control, añadir un lugar inicial y otro final.

1.  $Y_{L1} = \{p(A1, B1) = \{[a], \{b, e\}\}, p(A1, B2) = \{[a], \{c, e\}\}, p(B1, D1) = \{[e], \{d\}\}, p(B1, D1) = \{[b, e], \{d\}\}, P(B2, D1) = \{[c, e], \{d\}\}, \}$

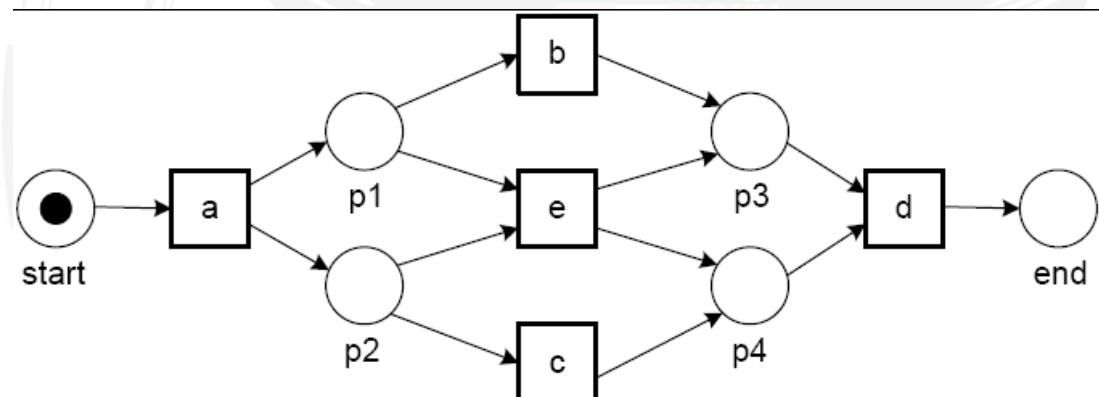


© Wil van der Aalst & TU/e (use only with permission & acknowledgements)

7. Determinar la relación de flujo: conectar cada plaza con cada elemento de su conjunto de transiciones de entrada y con cada elemento de su conjunto de transiciones de salida. Trazar un arco desde la plaza inicial a cada transición inicial (conjunto I) y desde cada transición final a la plaza final.



8. Devolver  $\alpha(L) = (P_L, T_L, F_L)$

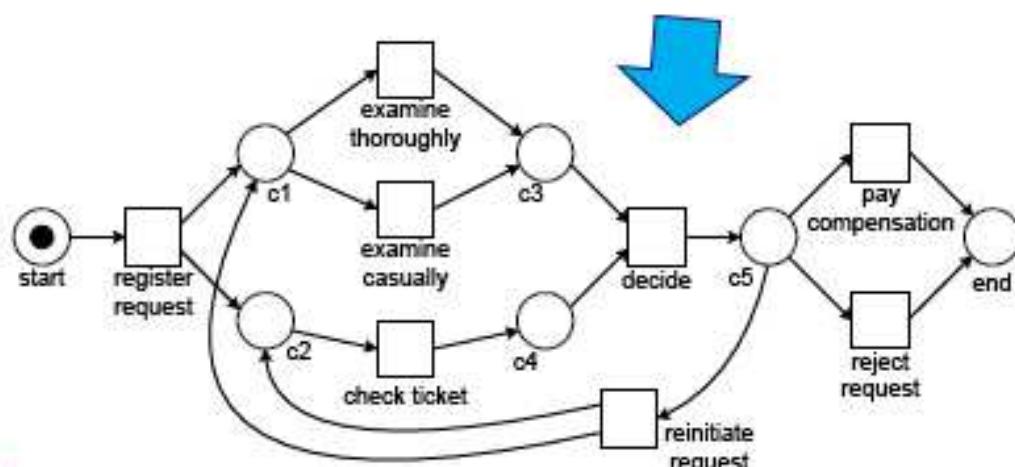


# Alpha algorithm

case id	event id	properties			
case id	event id	timestamp	activity	resource	cost
1	35654423	30-12-2010 11:00	register request	Pete	50
	35654424	31-12-2010 00:06	examine thoroughly	Sara	400
	35654425	05-01-2011 05:12	check ticket	Mike	100
	35654426	06-01-2011 01:18	decide	Sara	200
	35654427	07-01-2011 04:24	reject request	Pete	200
2	35654483	30-12-2010 11:32	register request	Mike	50
	35654485	30-12-2010 12:12	check ticket	Mike	100
	35654487	30-12-2010 14:16	examine casually	Pete	400
	35654488	05-01-2011 01:22	decide	Sara	200
	35654489	06-01-2011 12:06	pay compensation	Ellen	200
3	35654521	30-12-2010 14:32	register request	Pete	50
	35654522	30-12-2010 15:36	examine casually	Mike	400
	35654524	30-12-2010 16:34	check ticket	Ellen	100
	35654525	06-01-2011 09:18	decide	Sara	200
	35654526	06-01-2011 12:18	register request	Sara	200
	35654527	06-01-2011 13:06	examine thoroughly	Sara	400
	35654530	08-01-2011 01:43	check ticket	Pete	100
	35654531	09-01-2011 09:53	decide	Sara	200
	35654533	15-01-2011 00:43	pay compensation	Ellen	200
4	35654641	06-01-2011 15:02	register request	Pete	50
	35654643	07-01-2011 02:06	check ticket	Mike	100
	35654644	08-01-2011 04:43	examine thoroughly	Sara	400
	35654645	09-01-2011 02:02	decide	Sara	200
	35654647	12-01-2011 15:44	reject request	Ellen	200
5	35654711	06-01-2011 09:02	register request	Ellen	50
	35654712	07-01-2011 03:38	examine casually	Mike	400
	35654714	08-01-2011 11:22	check ticket	Pete	100
	35654715	10-01-2011 03:28	decide	Sara	200
	35654716	11-01-2011 06:18	register request	Sara	200
	35654718	14-01-2011 04:33	check ticket	Ellen	100
	35654719	16-01-2011 03:56	examine casually	Mike	400
	35654720	19-01-2011 01:38	decide	Sara	200
	35654721	20-01-2011 02:48	register request	Sara	200
	35654722	21-01-2011 09:06	examine casually	Sara	400
	35654724	21-01-2011 01:34	check ticket	Pete	100
	35654725	23-01-2011 03:32	decide	Sara	200
	35654726	24-01-2011 14:56	reject request	Mike	200
6	35654871	06-01-2011 05:02	register request	Mike	50
	35654872	06-01-2011 06:06	examine casually	Ellen	400
	35654874	07-01-2011 06:22	check ticket	Mike	100
	35654875	07-01-2011 06:52	decide	Sara	200
	35654877	16-01-2011 11:47	pay compensation	Mike	200

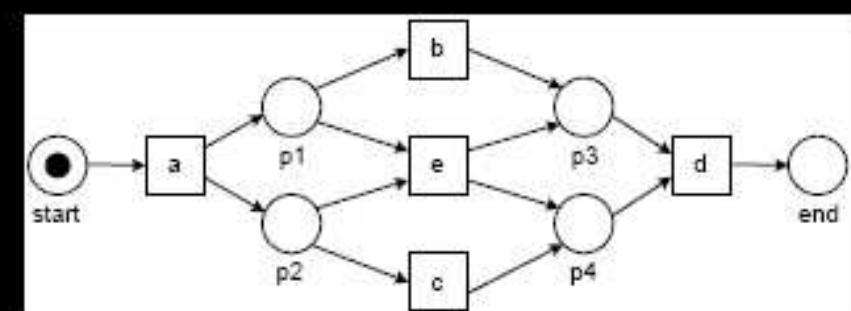


case id	trace
1	$\langle a, b, d, e, h \rangle$
2	$\langle a, d, c, e, g \rangle$
3	$\langle a, c, d, e, f, b, d, e, g \rangle$
4	$\langle a, d, b, e, h \rangle$
5	$\langle a, c, d, e, f, d, c, e, f, c, d, e, h \rangle$
6	$\langle a, c, d, e, g \rangle$



$$L_1 = [\langle a, b, c, d \rangle^3, \langle a, c, b, d \rangle^2, \langle a, e, d \rangle]$$

	$a$	$b$	$c$	$d$	$e$
$a$	$\#_{L_1}$	$\rightarrow_{L_1}$	$\rightarrow_{L_1}$	$\#_{L_1}$	$\rightarrow_{L_1}$
$b$	$\leftarrow_{L_1}$	$\#_{L_1}$	$\parallel_{L_1}$	$\rightarrow_{L_1}$	$\#_{L_1}$
$c$	$\leftarrow_{L_1}$	$\parallel_{L_1}$	$\#_{L_1}$	$\rightarrow_{L_1}$	$\#_{L_1}$
$d$	$\#_{L_1}$	$\leftarrow_{L_1}$	$\leftarrow_{L_1}$	$\#_{L_1}$	$\leftarrow_{L_1}$
$e$	$\leftarrow_{L_1}$	$\#_{L_1}$	$\#_{L_1}$	$\rightarrow_{L_1}$	$\#_{L_1}$



$$X_{L_1} = \{(\{a\}, \{b\}), (\{a\}, \{c\}), (\{a\}, \{e\}), (\{a\}, \{b, e\}), (\{a\}, \{c, e\}), (\{b\}, \{d\}), (\{c\}, \{d\}), (\{e\}, \{d\}), (\{b, e\}, \{d\}), (\{c, e\}, \{d\})\}$$

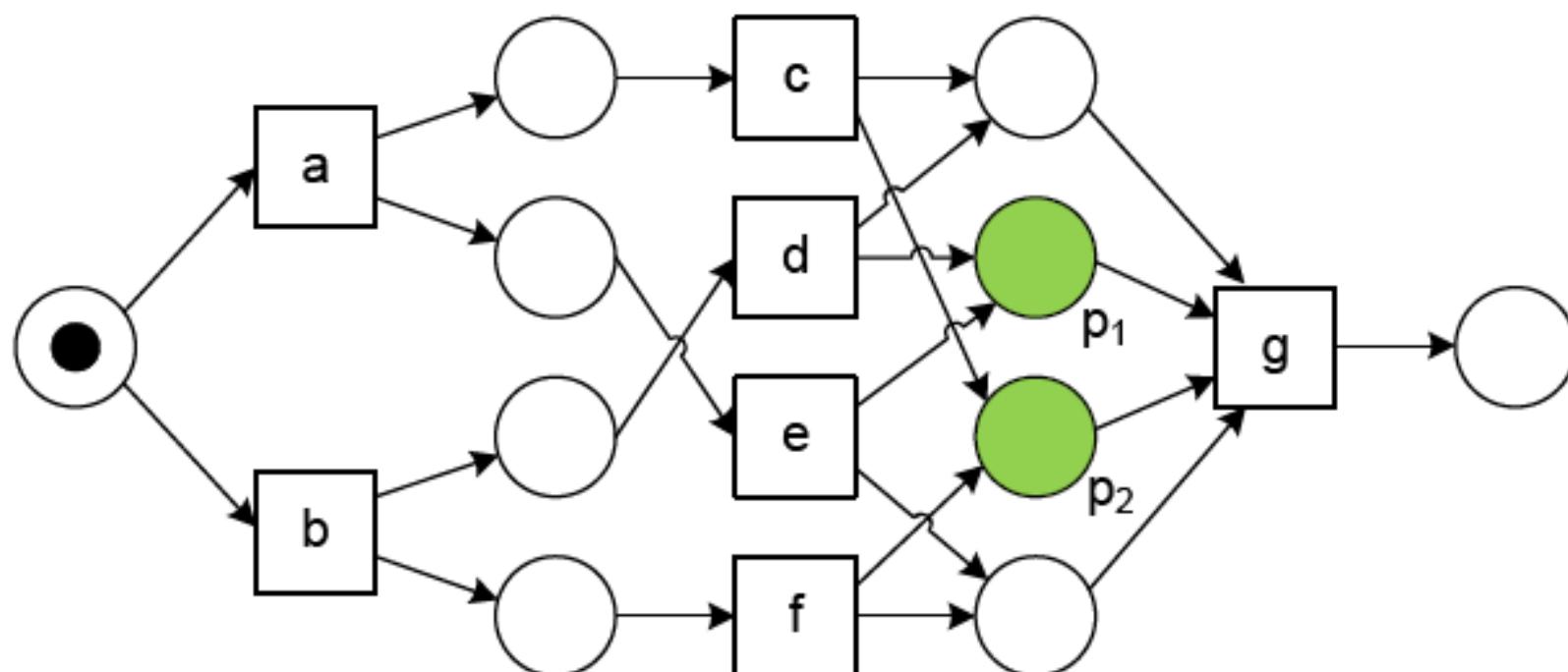
$$Y_{L_1} = \{(\{a\}, \{b, e\}), (\{a\}, \{c, e\}), (\{b, e\}, \{d\}), (\{c, e\}, \{d\})\}$$

- Una aproximación básica a minería de procesos (descubrimiento)
- Tiene limitaciones
- Aun así, ilustra los aspectos claves del descubrimiento de procesos:
  - Puede detectar patrones de control más comunes (secuencia, bucles, concurrencia, choices) con una aproximación simple.



Lugares implícitos: alpha puede duplicar lugares sin necesidad (redundancia).

$$L_6 = [\langle a, c, e, g \rangle^2, \langle a, e, c, g \rangle^3, \langle b, d, f, g \rangle^2, \langle b, f, d, g \rangle^4]$$



**$p_1$  and  $p_2$  are implicit places!**

- Lugares implícitos: alpha puede duplicar plazas.
- No puede descubrir repeticiones de una misma actividad (bucles de longitud 1)

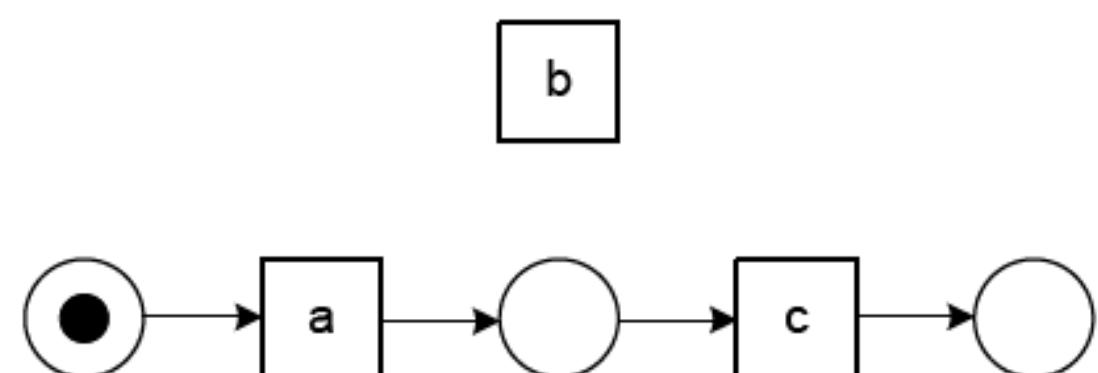
$$L_7 = [\langle a, c \rangle^2, \langle a, b, c \rangle^3, \langle a, b, b, c \rangle^2, \langle a, b, b, b, b, c \rangle^1]$$

a>b  
a>c  
b>b  
b>c

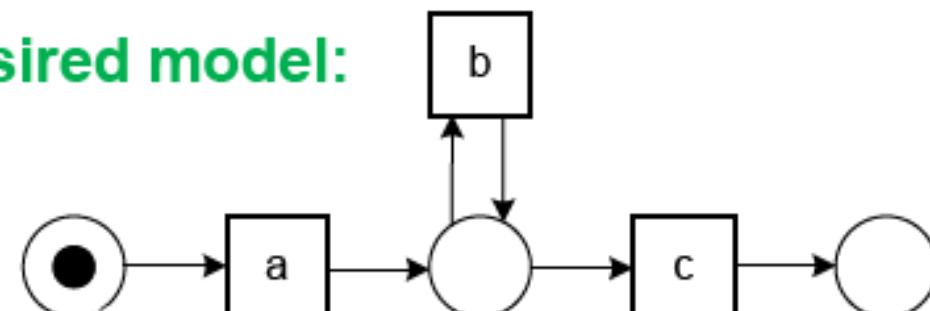
a→b  
a→c  
b→c

b||b

a#a  
c#c  
...



**desired model:**



Lugares implícitos: alpha puede duplicar plazas.

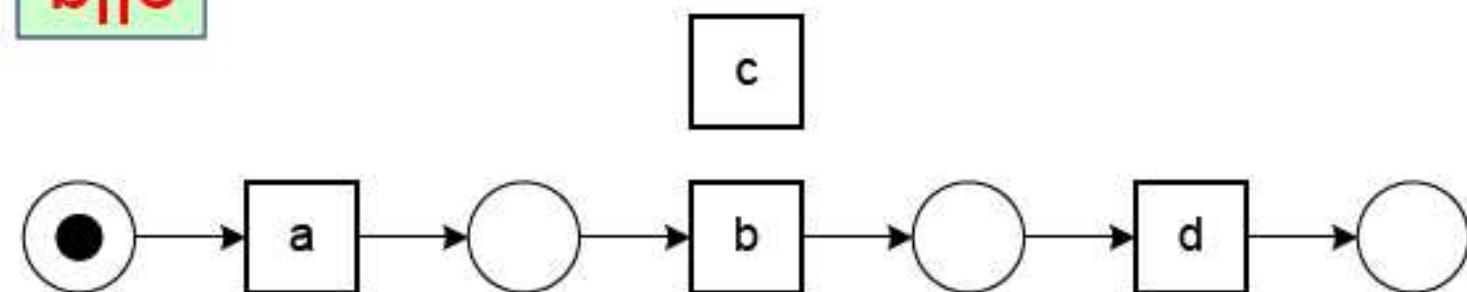
- No puede descubrir repeticiones de una misma actividad (bucles de longitud 1)
- Ni bucles de longitud 2.

$$L_8 = [\langle a, b, d \rangle^3, \langle a, b, c, b, d \rangle^2, \langle a, b, c, b, c, b, d \rangle]$$

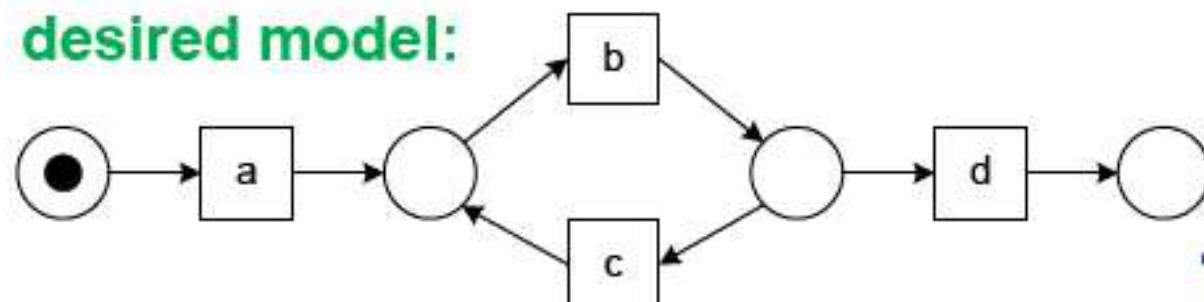
**a>b  
b>c  
b>d  
c>b**

**a→b  
b→d**

**b||c**

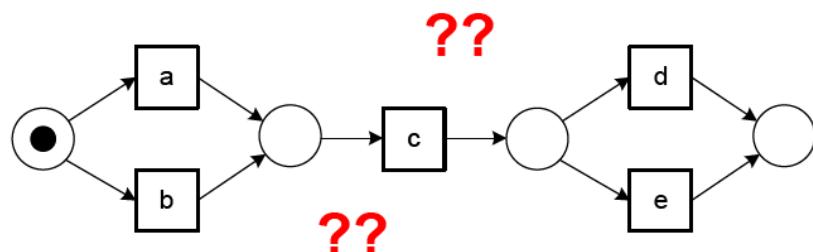


**desired model:**



- Lugares implícitos: alpha puede duplicar plazas.
- No puede descubrir repeticiones de una misma actividad (bucles de longitud 1)
- Ni bucles de longitud 2.
- Dependencias causales no locales

$$L_9 = [\langle a, c, d \rangle^{45}, \langle b, c, e \rangle^{42}]$$

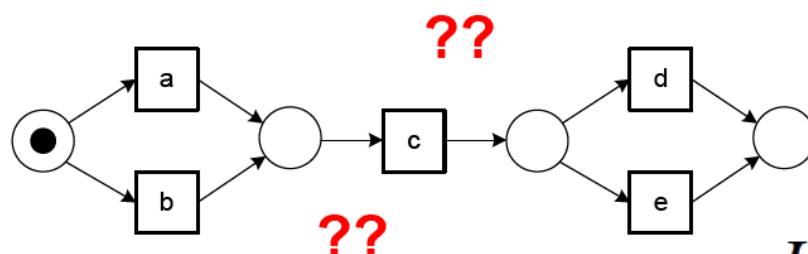


Observar que en esta PN puede dar lugar a las trazas bcd o ace, que no están inicialmente en el log

(e.d. dependencias causales entre actividades que están a una distancia mayor de dos, ed. que no se pueden detectar mediante la definición básica de dependencia causal).

- Lugares implícitos: alpha puede duplicar plazas.
- No puede descubrir repeticiones de una misma actividad (bucles de longitud 1)
- Ni bucles de longitud 2.
- Dependencias causales no locales

$$L_9 = [\langle a, c, d \rangle^{45}, \langle b, c, e \rangle^{42}]$$

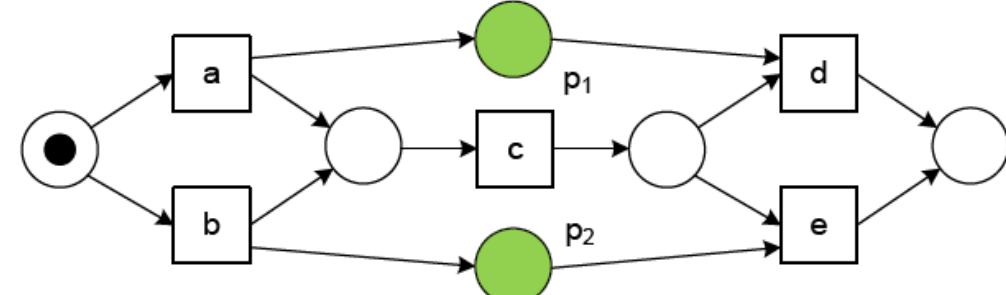


(e.d. dependencias causales entre actividades que están a una distancia mayor de dos, ed. que no se pueden detectar mediante la definición básica de dependencia causal).

Esto se arreglaría poniendo lugares que obliguen a que d se ejecute después de que a y c hayan terminado (y análogo para e). pero  $\alpha$  no lo detecta, no descubre las plazas

$$L_9 = [\langle a, c, d \rangle^{45}, \langle b, c, e \rangle^{42}]$$

Observar que en esta PN puede dar lugar a las trazas bcd o ace, que no están inicialmente en el log



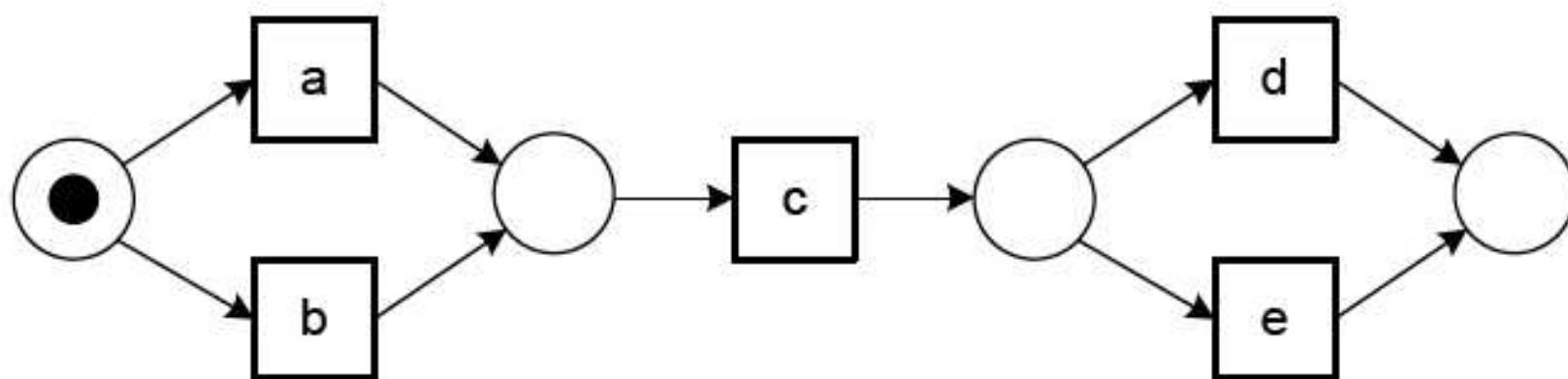
**p<sub>1</sub> and p<sub>2</sub> are not discovered!**

Lugares implícitos: alpha puede duplicar plazas.

- No puede descubrir repeticiones de una misma actividad (bucles de longitud 1)
- Ni bucles de longitud 2.
- Dependencias causales no locales
- Descubrir un mismo modelo para dos logs distintos

$$L_9 = [\langle a, c, d \rangle^{45}, \langle b, c, e \rangle^{42}]$$

$$L_4 = [\langle a, c, d \rangle^{45}, \langle b, c, d \rangle^{42}, \langle a, c, e \rangle^{38}, \langle b, c, e \rangle^{22}]$$

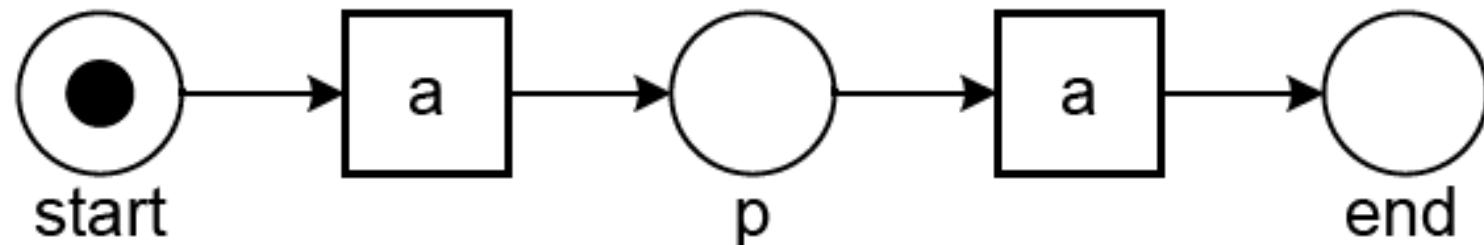


Lugares implícitos: alpha puede duplicar plazas.

- No puede descubrir repeticiones de una misma actividad (bucles de longitud 1)
- Ni bucles de longitud 2.
- Dependencias causales no locales  
Descubrir un mismo modelo para dos logs distintos
- Representational bias (sesgos del modelo de representación)

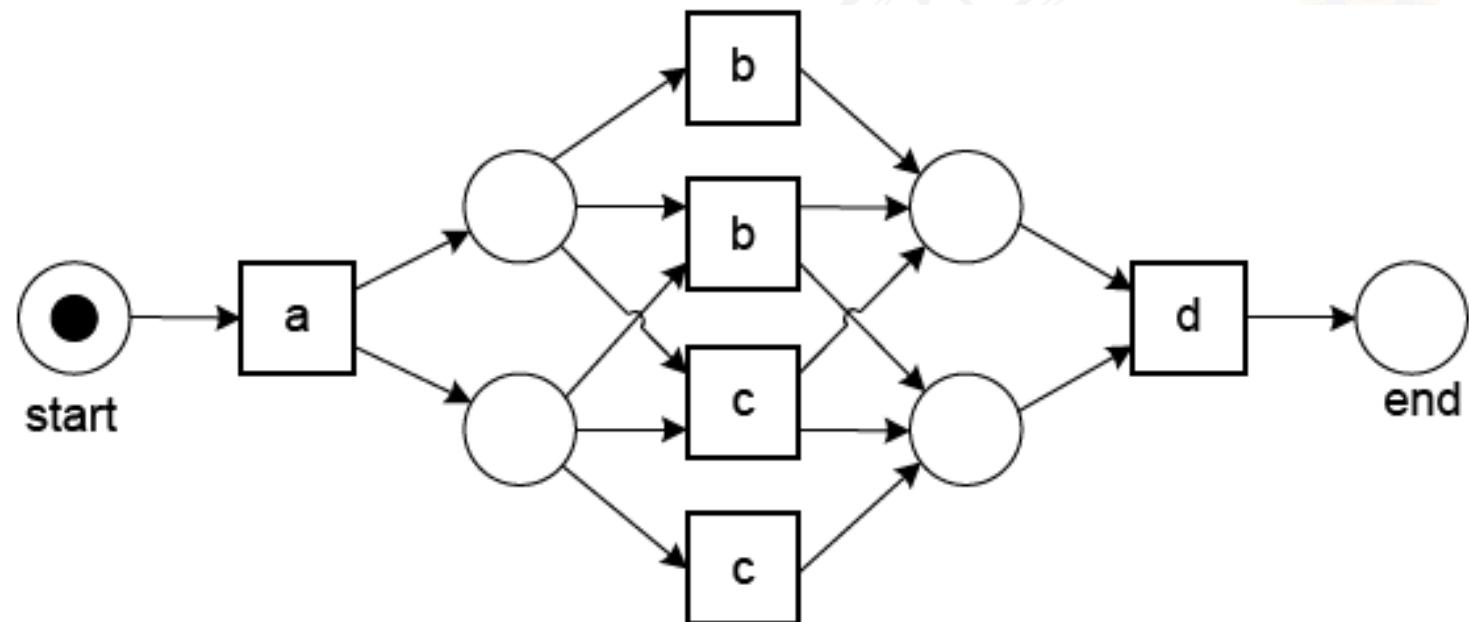
incapaz de descubrir un modelo con 2 transiciones (actividades) iguales, una seguida de la otra. Como WF-nets no pueden representar esta relación, La representación usada por  $\alpha$  no permitirá nunca descubrir este tipo de procesos

$$L_{10} = [\langle a, a \rangle^{55}]$$



- Lugares implícitos: alpha puede duplicar plazas.
- No puede descubrir repeticiones de una misma actividad (bucles de longitud 1)
- Ni bucles de longitud 2.
- Dependencias causales no locales  
Descubrir un mismo modelo para dos logs distintos
- Representational bias (sesgos del modelo de representación)

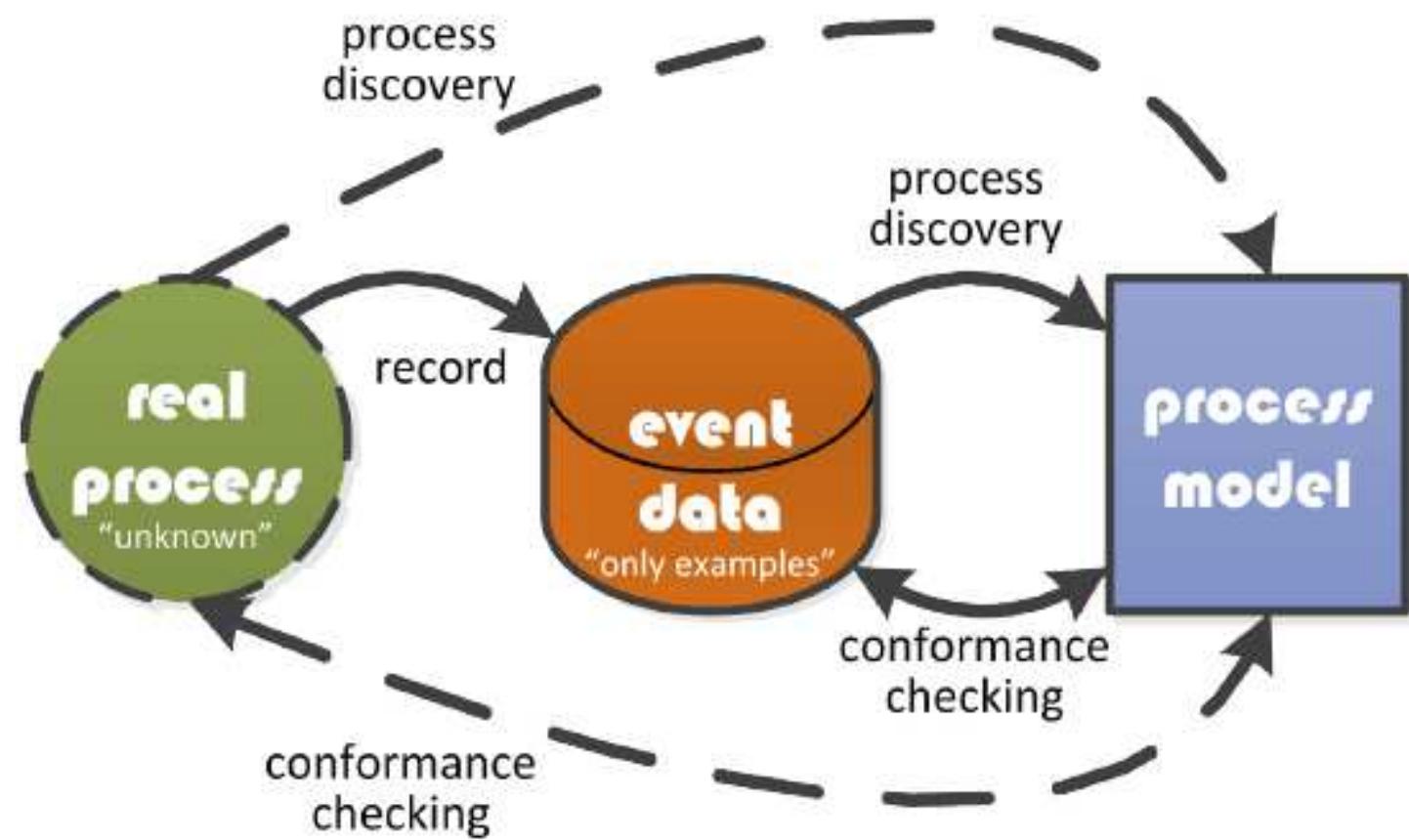
Esta WF-net se corresponde con un patrón de proceso muy común OR-split/join, no puede aprenderse con WF-net ni con alpha.

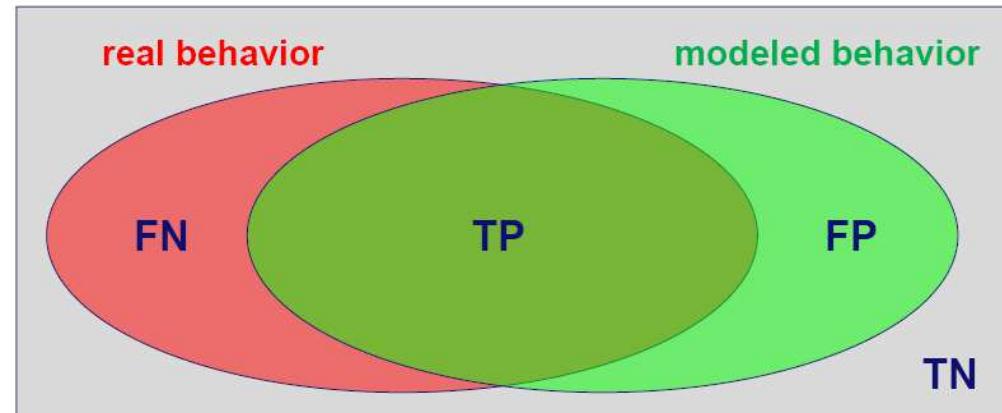


- Lugares implícitos: alpha puede duplicar plazas.
- No puede descubrir repeticiones de una misma actividad (bucles de longitud 1)
- Ni bucles de longitud 2.
- Dependencias causales no locales  
Descubrir un mismo modelo para dos logs distintos
- Representational bias (sesgos del modelo de representación)
- **Puede descubrir modelos que no son correctos.**

- Desafío: ruido e incompletitud.
  - Para descubrir un modelo de proceso aceptable hay que asumir que el log de eventos contiene una muestra representativa del comportamiento a aprender.
- Ruido
  - El log contiene trazas raras o infrecuentes no representativas del comportamiento típico esperado.
- Incompletitud:
  - El log contiene demasiados pocos eventos para descubrir algunas de las estructuras de control de flujo subyacentes en el modelo.

¿El modelo de proceso obtenido es un reflejo correcto del proceso real?





		predicted class	
		+	-
actual class	+	TP	FN
	-	FP	TN

- FN (Falsos Negativos): trazas que existen en la realidad pero que no pueden reproducirse con el modelo.
- TP(Ciertos Positivos): trazas que existen en la realidad y que se pueden reproducir con el modelo.
- FP(Falsos Positivos): trazas que no existen en la realidad y que se pueden reproducir con el modelo.
- TN(Ciertos Negativos): trazas que no existen en la realidad y que no se pueden reproducir con el modelo.

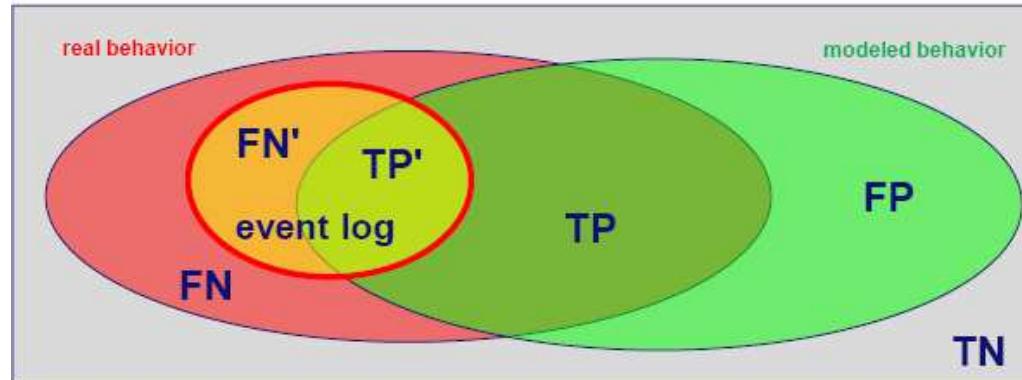
$$\text{recall} = \frac{TP}{TP + FN}$$

$$\text{precision} = \frac{TP}{TP + FP}$$

El log de eventos contiene típicamente una fracción de trazas posibles

~~$$\text{recall} = \frac{TP}{TP + FN}$$~~

~~$$\text{precision} = \frac{TP}{TP + FP}$$~~

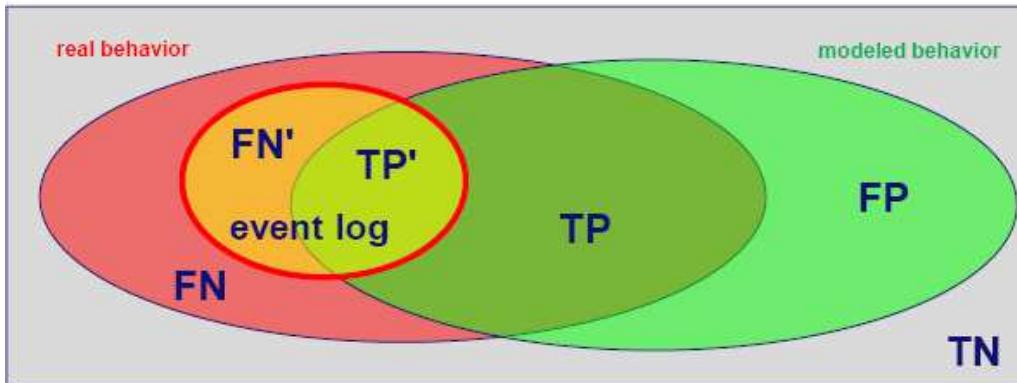


		predicted class	
		+	-
actual class	+	TP	FN
	-	FP	TN

- Falsos Negativos: trazas que existen en la realidad pero que no pueden reproducirse con el modelo.
- Ciertos Positivos: trazas que existen en la realidad y que se pueden reproducir con el modelo.
- ~~Falsos Positivos: trazas que no existen en la realidad y que se pueden reproducir con el modelo.~~
- ~~Ciertos Negativos: trazas que no existen en la realidad y que no se pueden reproducir con el modelo.~~

$$\text{replay\_fitness} = \frac{TP'}{TP' + FN'}$$

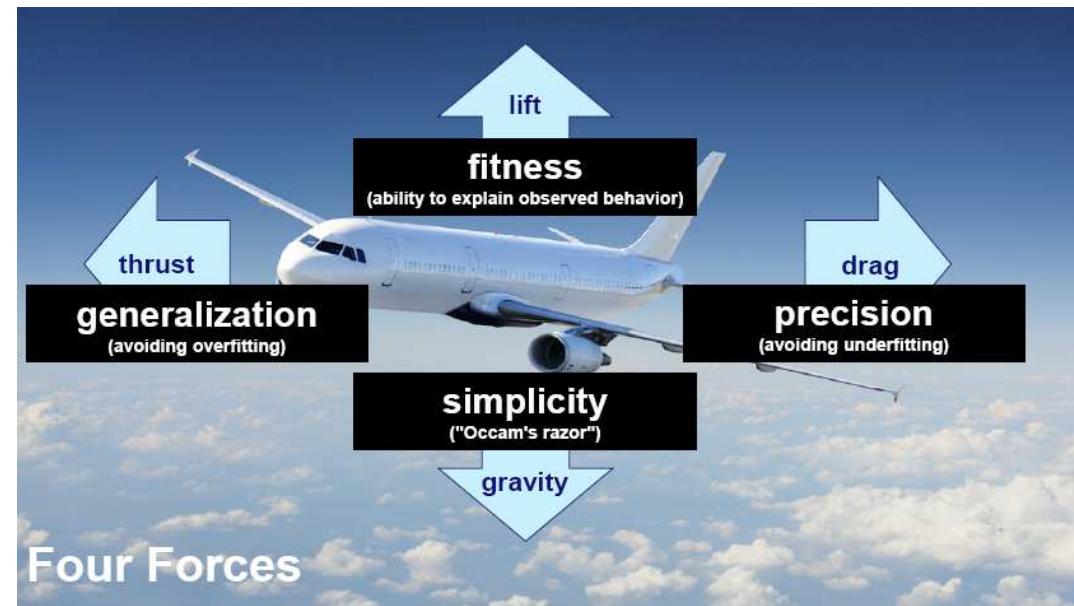
El log de eventos contiene típicamente una fracción de trazas posibles



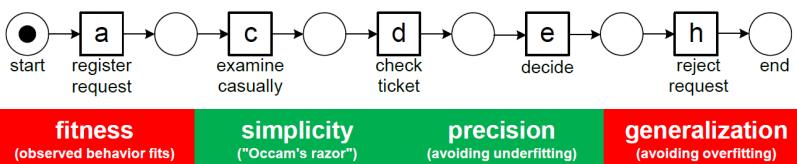
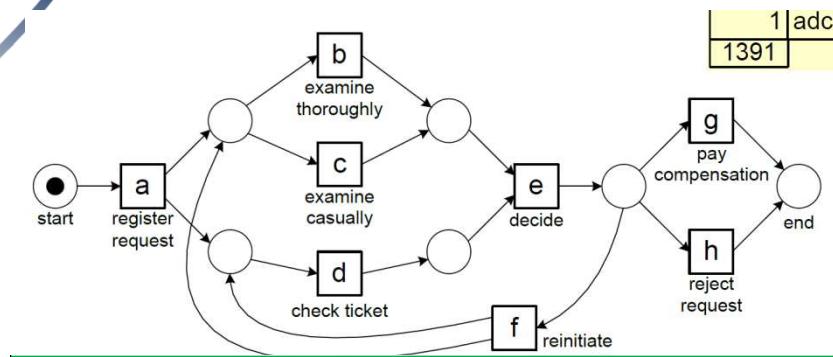
$$\text{replay\_fitness} = \frac{TP'}{TP' + FN'}$$

- Falsos Negativos: trazas que existen en la realidad pero que no pueden reproducirse con el modelo.
- Ciertos Positivos: trazas que existen en la realidad y que se pueden reproducir con el modelo.
- ~~Falsos Positivos: trazas que no existen en la realidad y que se pueden reproducir con el modelo.~~
- ~~Ciertos Negativos: trazas que no existen en la realidad y que no se pueden reproducir con el modelo.~~

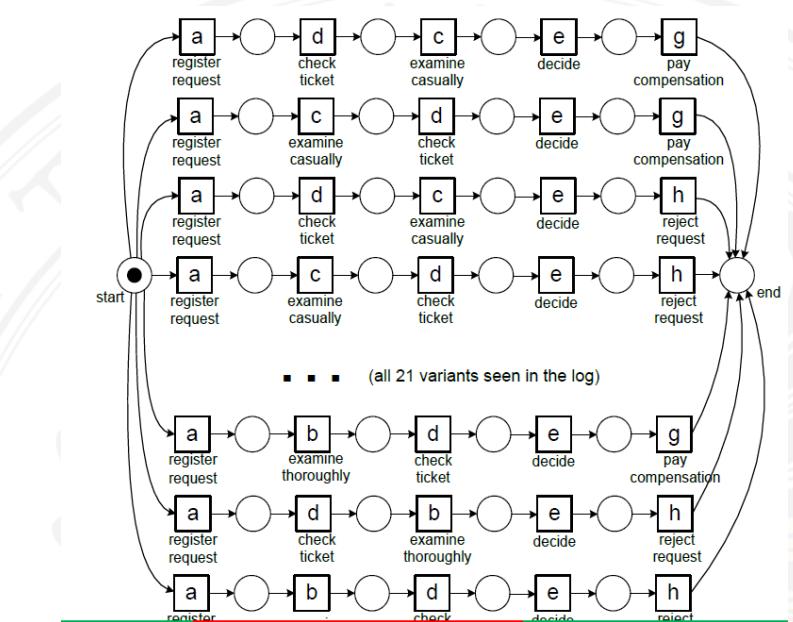
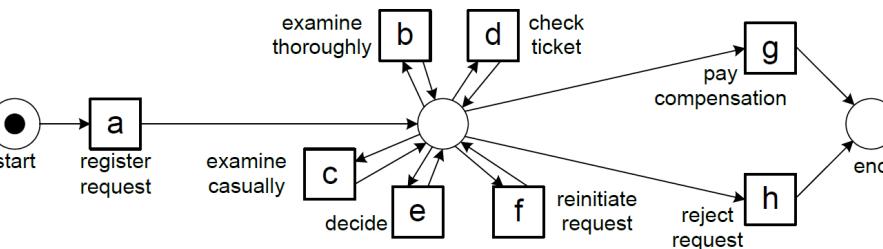
- Desafíos:
  - No hay ejemplos negativos, e.d., (no puedo ver lo que no ocurre)
  - El log contiene una fracción de trazas posibles.
  - Trazas que “casi se ajustan” o que tienen “ajuste pobre”.
  - Si hay bucles a menudo tendremos infinitas trazas posibles.
  - Ley de Murphy para PM: cualquier cosa es posible, las probabilidades cuentan.



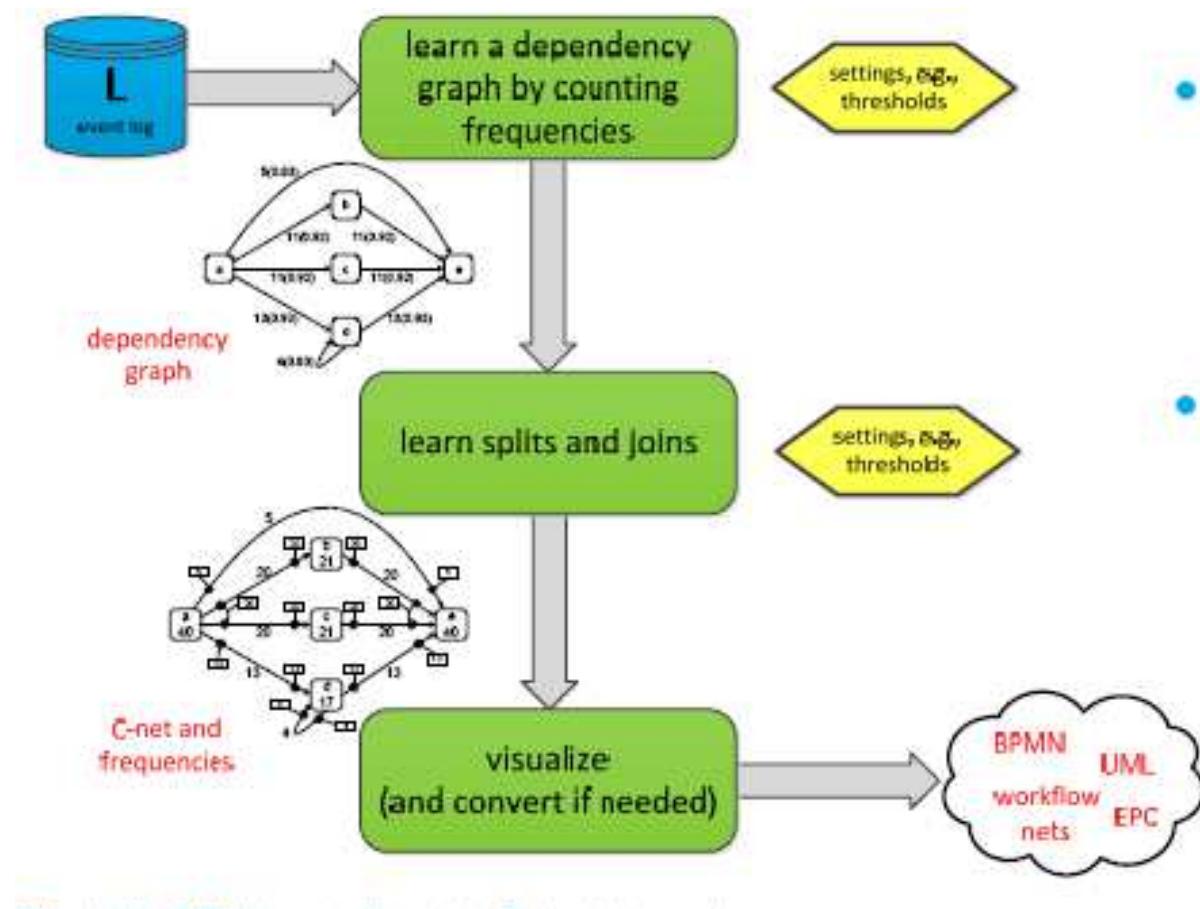
- ¿Qué criterios puedo considerar para establecer una medida de calidad (aunque sea cualitativa) del algoritmo de process mining?
  - Fitness: capacidad de explicar el comportamiento observado (¿explica todas las trazas del log?).
  - Simplicity: capacidad de explicarlo de una manera simple.
    - Navaja de Ocam: ante dos teorías que explican los mismos hechos observados, es mejor optar por la más simple.
  - Precisión: capacidad de explicar las trazas de forma precisa, o evitar holgura (underfitting) (¿explica más trazas de las observadas?)
  - Generalización: capacidad de explicar trazas aun no observadas, o evitar sobreajuste (overfitting) (¿puede explicar trazas no observadas reales, falsos negativos?)



#	trace
455	acdeh
191	abdeg
177	adceh
144	abdeh
111	acdeg
82	adceg
56	adbeh
47	acdefdbeh
38	adbeg
33	acdefbdeh
14	acdefbdieg
11	acdefdbeg
9	adcefcdbeh
8	adcefdbeh
5	adcefbdeg
3	acdefbdefdbeg
2	adcefdieg
2	adcefbdefbdeg
1	adcefdbefbdbeh
1	adbefbdefdbeg
1	adcefdbefcdefdbeg
391	



Los algoritmos de process mining heurístico consideran frecuencias de eventos y secuencias cuando construyen el modelo de proceso. Idea: caminos infrecuentes no deberían incorporarse en el modelo. El uso de C-nets y de frecuencias consiguen que estas aproximaciones sean más robustas (descubren modelos de proceso más ricos).



$$L = [\langle a, e \rangle^5, \langle a, b, c, e \rangle^{10}, \langle a, c, b, e \rangle^{10}, \langle a, b, e \rangle^1, \langle a, c, e \rangle^1, \\ \langle a, d, e \rangle^{10}, \langle a, d, d, e \rangle^2, \langle a, d, d, d, e \rangle^1]$$

- Sucesión directa:  $x > y$  sií para algún caso  $x$  está seguida directamente por  $y$ .
- Causalidad:  $x \rightarrow y$  sií  $x > y$  and  $\text{not}(y > x)$

Número de veces que  $a$  es seguida directamente por  $b$

$$|a >_L b| = \sum_{\sigma \in L} L(\sigma) \times |\{1 \leq i < |\sigma| \mid \sigma(i) = a \wedge \sigma(i+1) = b\}|$$

direct succession

### dependency measure

$|a \Rightarrow_L b|$  is the value of the dependency relation between  $a$  and  $b$ :

$$|a \Rightarrow_L b| = \begin{cases} \frac{|a >_L b| - |b >_L a|}{|a >_L b| + |b >_L a| + 1} & \text{if } a \neq b \\ \frac{|a >_L a|}{|a >_L a| + 1} & \text{if } a = b \end{cases}$$

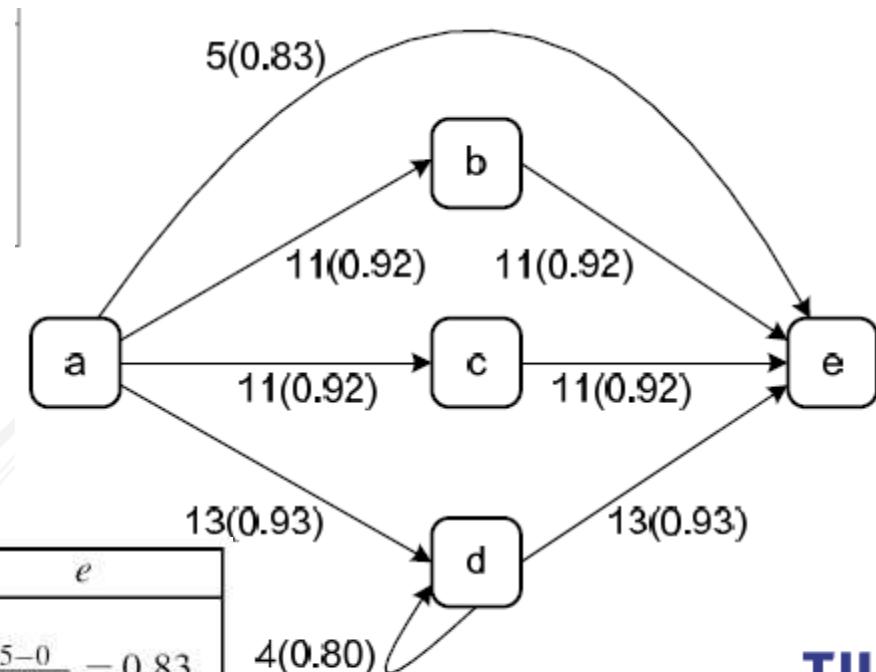
Valor de la relación de dependencia causal entre  $a$  y  $b$ .

Caso particular si ambas son iguales.

En  $[-1, 1]$

$$L = [\langle a, e \rangle^5, \langle a, b, c, e \rangle^{10}, \langle a, c, b, e \rangle^{10}, \langle a, b, e \rangle^1, \langle a, c, e \rangle^1, \\ \langle a, d, e \rangle^{10}, \langle a, d, d, e \rangle^2, \langle a, d, d, d, e \rangle^1]$$

$ >_L $	$a$	$b$	$c$	$d$	$e$
$a$	0	11	11	13	5
$b$	0	0	10	0	11
$c$	0	10	0	0	11
$d$	0	0	0	4	13
$e$	0	0	0	0	0



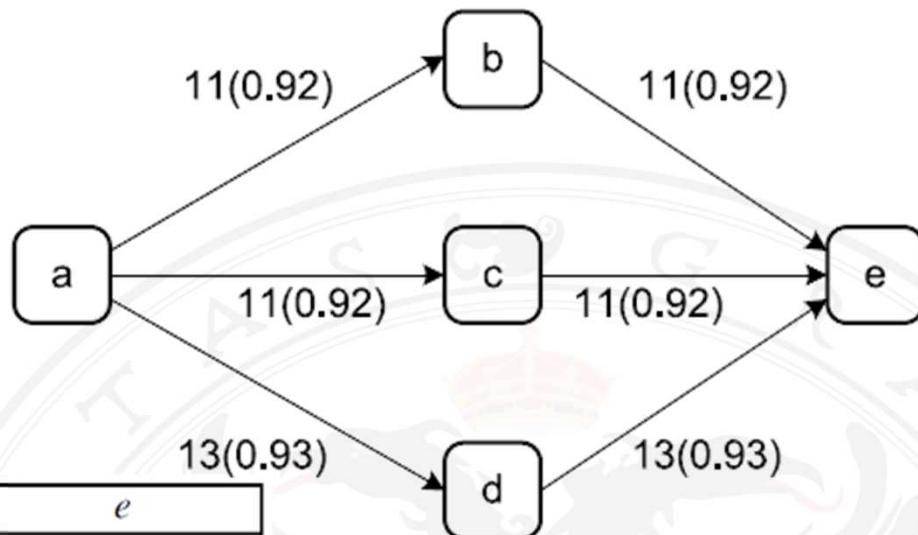
$ >_L $	$a$	$b$	$c$	$d$	$e$
$a$	$\frac{0}{0+1} = 0$	$\frac{11-0}{11+0+1} = 0.92$	$\frac{11-0}{11+0+1} = 0.92$	$\frac{13-0}{13+0+1} = 0.93$	$\frac{5-0}{5+0+1} = 0.83$
$b$	$\frac{0-11}{0+11+1} = -0.92$	$\frac{0}{0+1} = 0$	$\frac{10-10}{10+10+1} = 0$	$\frac{0-0}{0+0+1} = 0$	$\frac{11-0}{11+0+1} = 0.92$
$c$	$\frac{0-11}{0+11+1} = -0.92$	$\frac{10-10}{10+10+1} = 0$	$\frac{0}{0+1} = 0$	$\frac{0-0}{0+0+1} = 0$	$\frac{11-0}{11+0+1} = 0.92$
$d$	$\frac{0-13}{0+13+1} = -0.93$	$\frac{0-0}{0+0+1} = 0$	$\frac{0-0}{0+0+1} = 0$	$\frac{4}{4+1} = 0.80$	$\frac{13-0}{13+0+1} = 0.93$
$e$	$\frac{0-5}{0+5+1} = -0.83$	$\frac{0-11}{0+11+1} = -0.92$	$\frac{0-11}{0+11+1} = -0.92$	$\frac{0-13}{0+13+1} = -0.93$	$\frac{0}{0+1} = 0$



- Podemos hacer uso de valores umbral para determinar la fortaleza de las relaciones

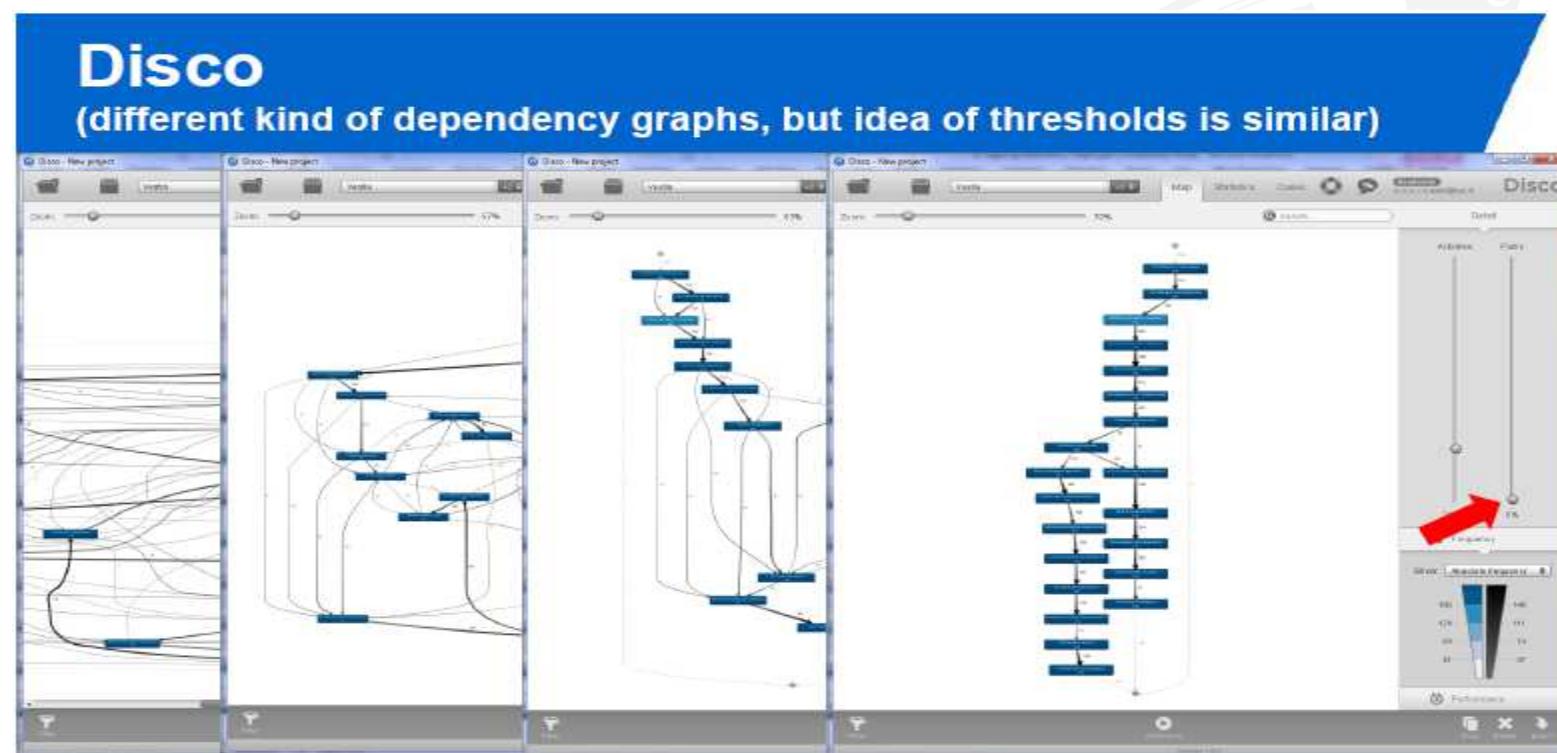
$$L = [\langle a, e \rangle^5, \langle a, b, c, e \rangle^{10}, \langle a, c, b, e \rangle^{10}, \langle a, b, e \rangle^1, \langle a, c, e \rangle^1, \\ \langle a, d, e \rangle^{10}, \langle a, d, d, e \rangle^2, \langle a, d, d, d, e \rangle^1]$$

$ >_L $	$a$	$b$	$c$	$d$	$e$
$a$	0	11	11	13	5
$b$	0	0	10	0	11
$c$	0	10	0	0	11
$d$	0	0	0	4	13
$e$	0	0	0	0	0



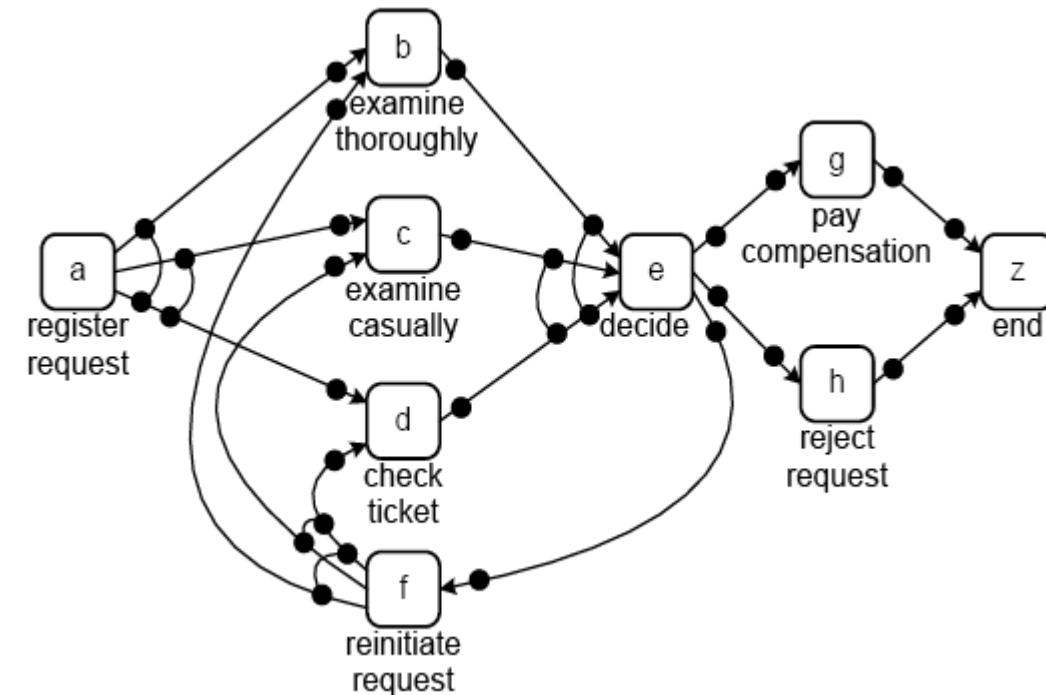
$ \Rightarrow_L $	$a$	$b$	$c$	$d$	$e$
$a$	$\frac{0}{0+1} = 0$	$\frac{11-0}{11+0+1} = 0.92$	$\frac{11-0}{11+0+1} = 0.92$	$\frac{13-0}{13+0+1} = 0.93$	$\frac{5-0}{5+0+1} = 0.83$
$b$	$\frac{0-11}{0+11+1} = -0.92$	$\frac{0}{0+1} = 0$	$\frac{10-10}{10+10+1} = 0$	$\frac{0-0}{0+0+1} = 0$	$\frac{11-0}{11+0+1} = 0.92$
$c$	$\frac{0-11}{0+11+1} = -0.92$	$\frac{10-10}{10+10+1} = 0$	$\frac{0}{0+1} = 0$	$\frac{0-0}{0+0+1} = 0$	$\frac{11-0}{11+0+1} = 0.92$
$d$	$\frac{0-13}{0+13+1} = -0.93$	$\frac{0-0}{0+0+1} = 0$	$\frac{0-0}{0+0+1} = 0$	$\frac{4}{4+1} = 0.80$	$\frac{13-0}{13+0+1} = 0.93$
$e$	$\frac{0-5}{0+5+1} = -0.83$	$\frac{0-11}{0+11+1} = -0.92$	$\frac{0-11}{0+11+1} = -0.92$	$\frac{0-13}{0+13+1} = -0.93$	$\frac{0}{0+1} = 0$

- Como calcular el grafo de dependencias:
  1. Poner los umbrales para el valor mínimo de sucesiones directas y medida de dependencia.
  2. Contar sucesiones directas en la footprint.
  3. Calcular medidas de dependencia.
  4. Dibujar el grafo obtenido a partir de la footprint como matriz de adyacencia, considerando sólo arcos que se ajustan a los umbrales.



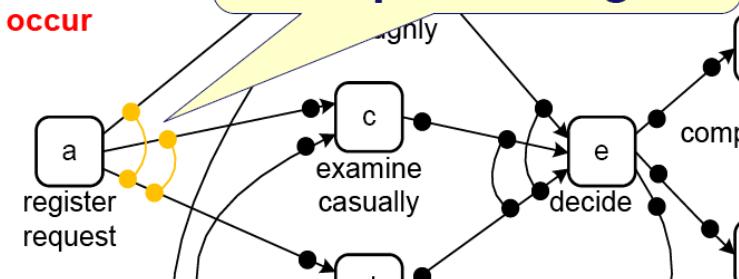
C-nets es una representación destinada a process mining.

- Es una extensión de un grafo de dependencias causales, inspirada en Pnets.
- Los arcos representan las dependencias causales, igual que un grafo de dependencias.
- Cada actividad tiene un conjunto de ligaduras de entrada y ligaduras de salida (input binding, output binding).
  - Cada ligadura puede incluir uno o más arcos.
- a tiene dos posibles output bindings
  - Es seguida por {b y d} o bien por {c y d}
- e tiene dos posibles input bindings
- e tiene tres posibles output bindings.



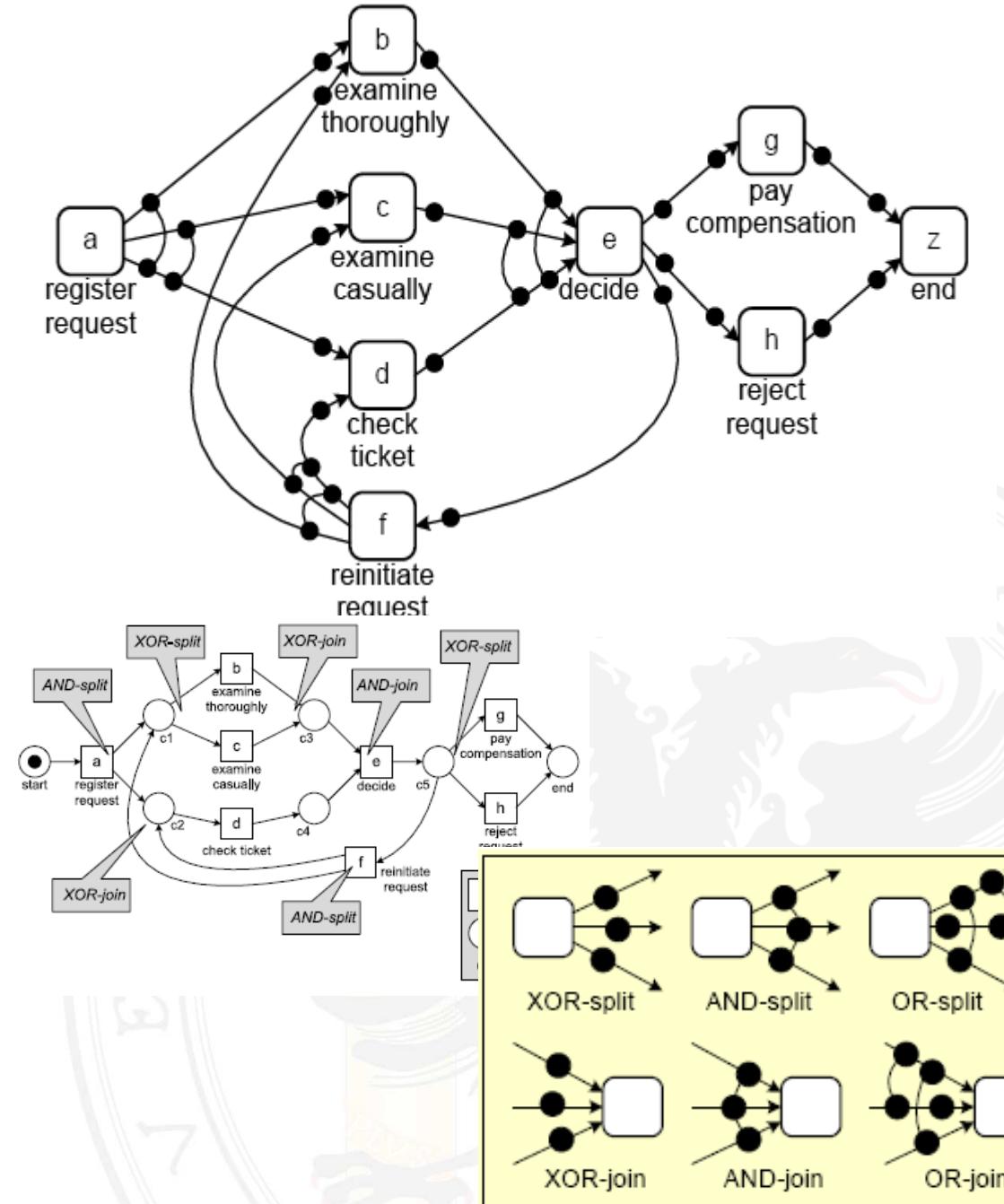
activity a is the start activity and will be the first to occur

activity a has two output bindings

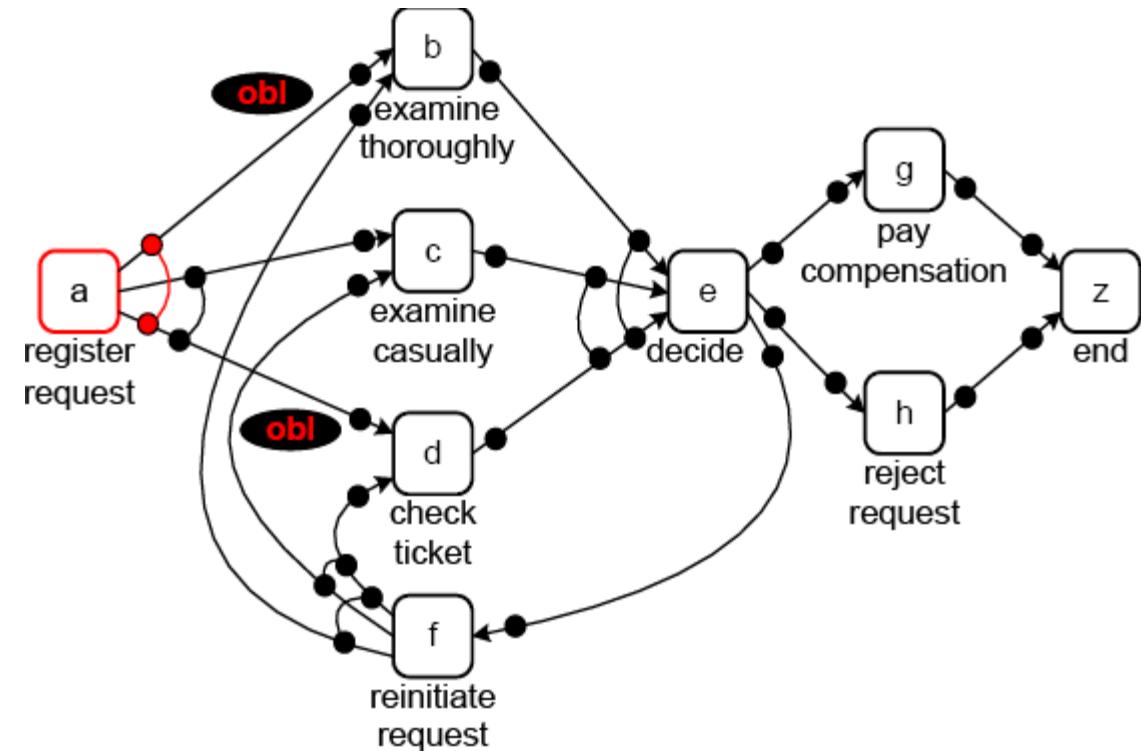


C-nets es una representación destinada a process mining.

- Es una extensión de un grafo de dependencias causales, inspirada en Pnets.
- Los arcos representan las dependencias causales, igual que un grafo de dependencias.
- Cada actividad tiene un conjunto de ligaduras de entrada y ligaduras de salida (input binding, output binding).
  - Cada ligadura puede incluir uno o más arcos.
- **Las ligaduras (bindings) permiten expresar de forma más precisa y simplificada mayor variedad de patrones de proceso comunes.**

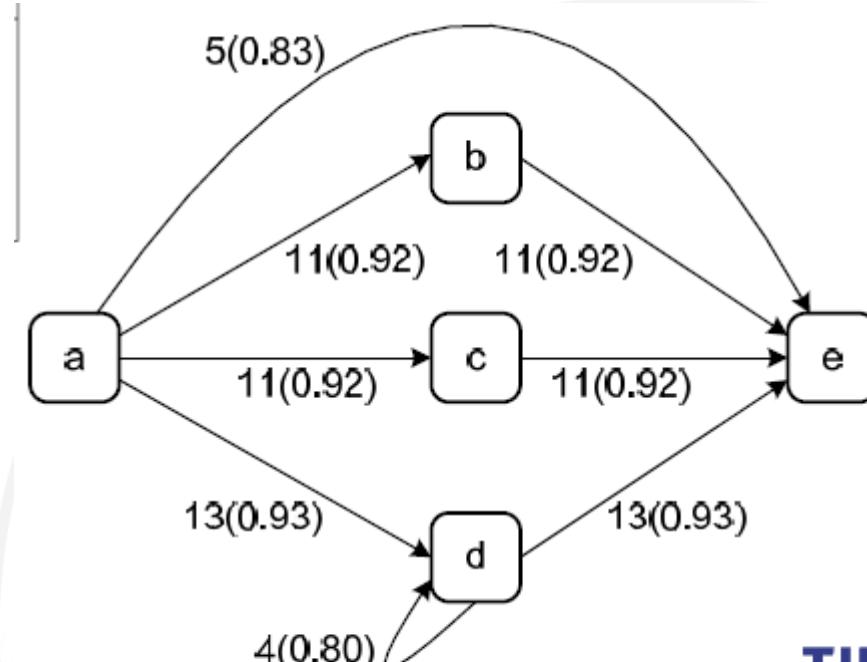


- Suministra una semántica para poder hacer replay, más que para ejecución.
- En lugar de tokens tenemos “obligations”.
- Por ejemplo, si nos llega para hacer replay la traza  $\{a,b,d,e,g,z\}$  detectará que a se ejecuta y activa el binding  $\{b,d\}$  creando dos obligations.
- Las obligations activan b y d
- ...



El objetivo de PM heurístico es generar una C-net a partir de un log, generando un Grafo de Dependencia Causal como paso intermedio. Los nodos y arcos de la C-net corresponden a los nodos y arcos del grafo, solo queda aprender los bindings.

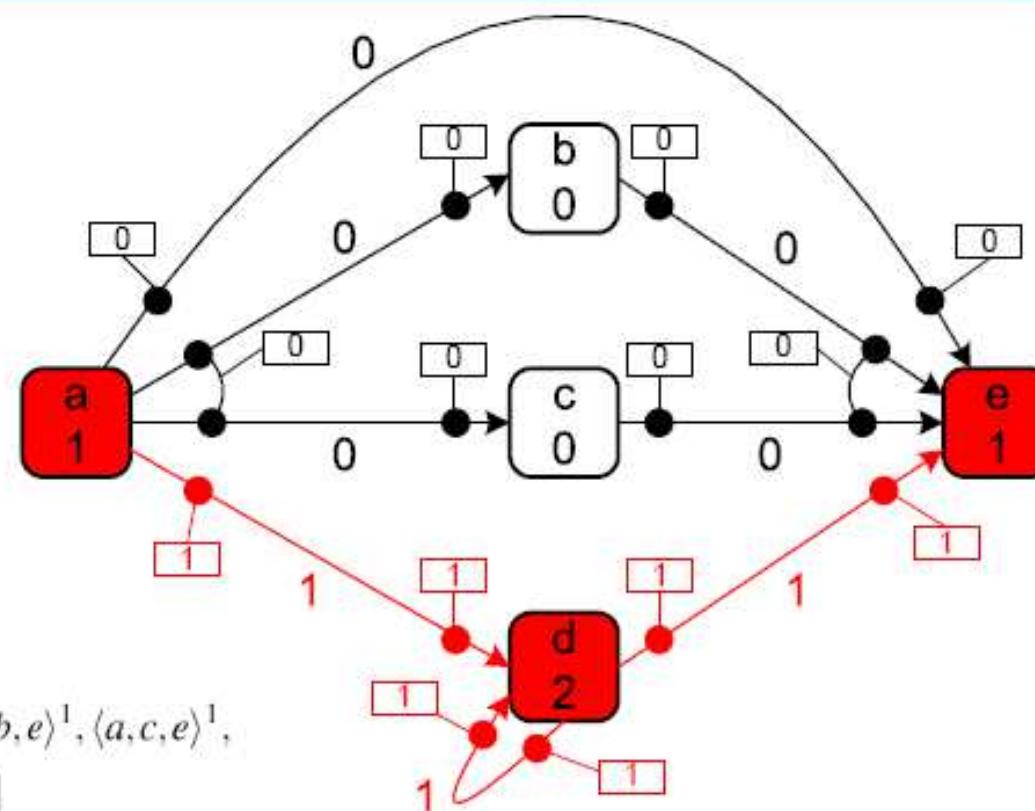
- Determinar los posibles bindings de entrada y salida para cada actividad.
- a tiene dependencias con  $\{b,c,d,e\}$ , por tanto hay  $2^4-1$  posibles bindings de salida  $\{\{b\}, \{c\}, \{d\}, \{e\}, \{b,c\}, \{b,d\}, \dots, \{b,c,d,e\}\}$ .
- b sólo tiene un posible binding de salida
- d tiene 3 bindings potenciales de salida, y 3 potenciales bindings de entrada.
- ...



TU/e

Si hay sólo un binding de entrada o salida, este es el que se toma. Para otros bindings, hay que seleccionar subconjuntos y esto se hace haciendo replay del log sobre el grafo de dependencias.

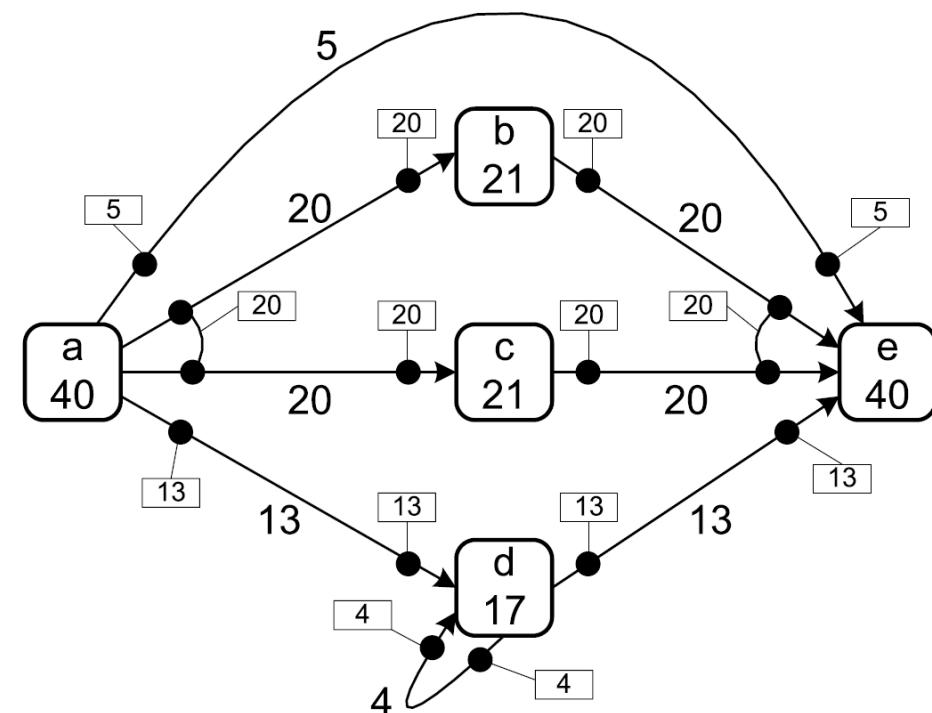
## Example path: adde



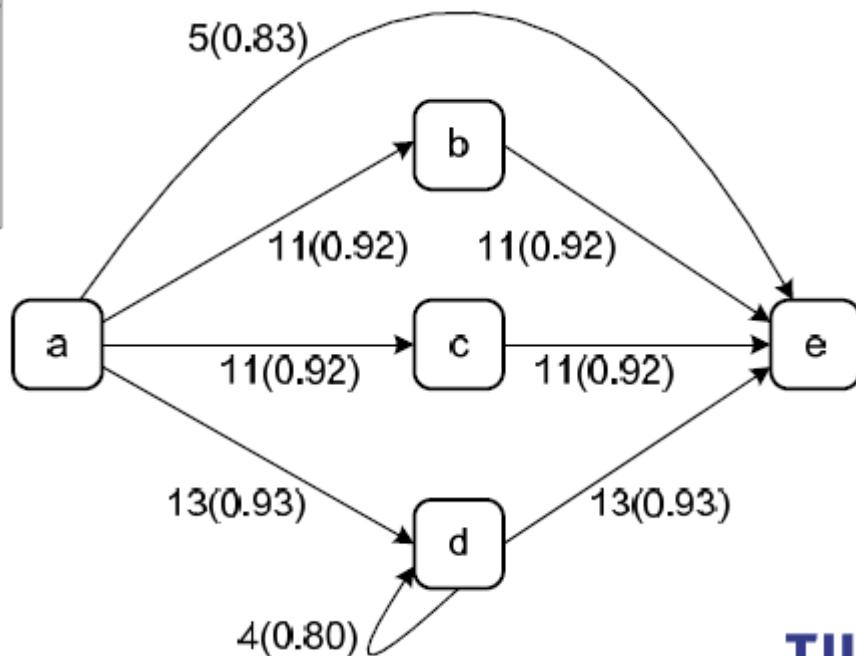
$$L = [\langle a, e \rangle^5, \langle a, b, c, e \rangle^{10}, \langle a, c, b, e \rangle^{10}, \langle a, b, e \rangle^1, \langle a, c, e \rangle^1, \\ \langle a, d, e \rangle^{10}, \langle a, d, d, e \rangle^2, \langle a, d, d, d, e \rangle^1]$$

Si hacemos un replay con secuencias que empiezan por a, que tiene 15 posibles bindings, obtenemos que a es seguida por e 5 veces, seguida por {b,c} 20 veces  $\langle abce \rangle \langle acbe \rangle$ , seguida por d 13 veces  $\langle ade \rangle \langle adde \rangle$  y una vez por c  $\langle ace \rangle$ . Supongamos que tenemos un umbral bajo el cual no se consideran output bindings infrecuentes, entonces de los 15 posibles, algunos están en el log pero no se incluyen, y en la C-net solo aparecen 3  $\{\{e\}, \{d\}, \{b, c\}\}$ .

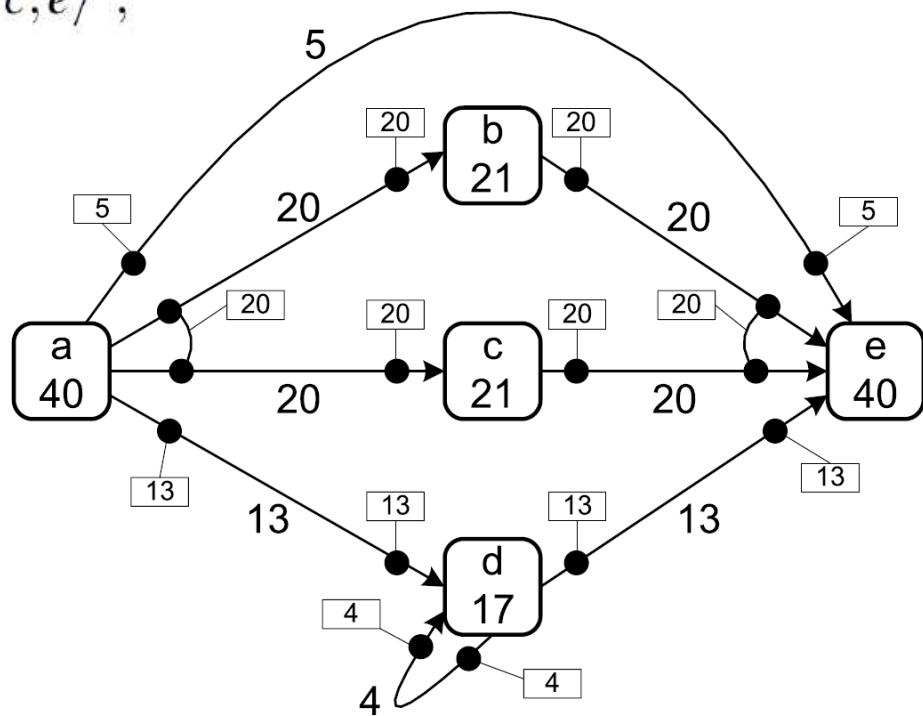
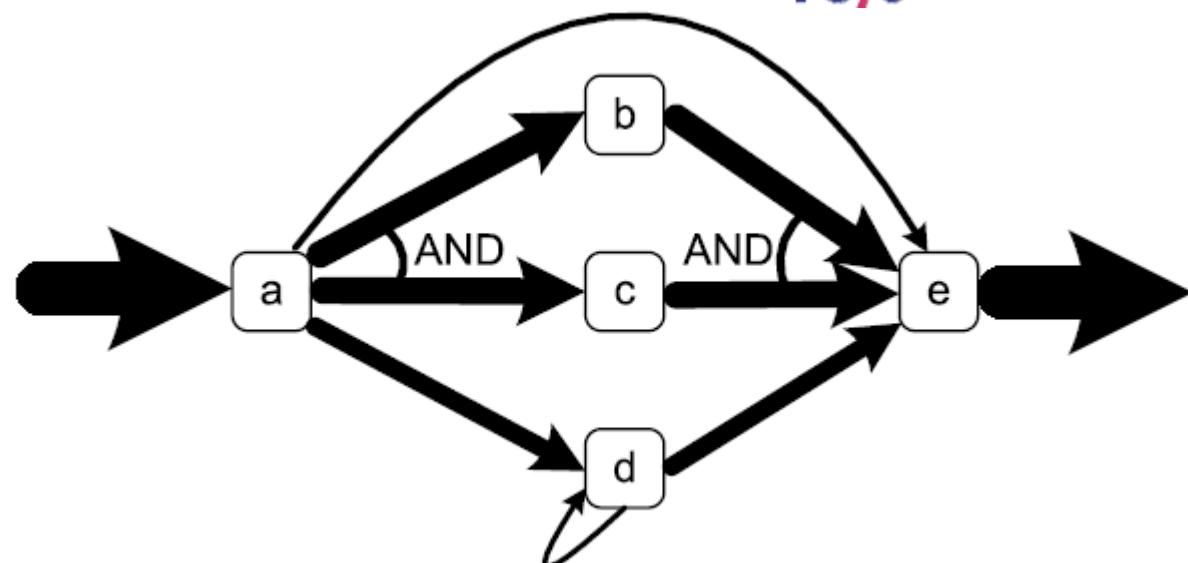
- Haciendo replay podemos obtener frecuencias de input y output bindings.
- Frecuencias mostradas:
  - Cada nodo muestra la frecuencia de su actividad correspondiente. (a = 40)
  - Cada arco tiene una frecuencia indicando cuántas veces ambas actividades están en un binding común (a,c) = 20.
  - Frecuencias de bindings de entrada y salida.
    - A -> {b,c} = 20



$$L = [\langle a, e \rangle^5, \langle a, b, c, e \rangle^{10}, \langle a, c, b, e \rangle^{10}, \langle a, b, e \rangle^1, \langle a, c, e \rangle^1, \\ \langle a, d, e \rangle^{10}, \langle a, d, d, e \rangle^2, \langle a, d, d, d, e \rangle^1]$$



**TU/e**



- Auditoría y cumplimiento
- Evaluación de algoritmos de process discovery
- Pasar de conformidad a especificación (software, servicios, etc).

