



Process Mining

Come estrarre conoscenza dai log dei sistemi informativi orientati ai processi

W.M.P. Van Der Aalst, A. Burattin, M. de Leoni, A. Guzzo,
F.M. Maggi, M. Montali

Le tecniche di process mining consentono di estrarre conoscenza dai log generati dai sistemi informativi e più specificamente di modellare, monitorare, e migliorare i processi presenti in un'ampia varietà di domini applicativi. Il crescente interesse nei confronti di questa disciplina è dovuto sia alla sempre maggiore disponibilità di dati che forniscono informazioni dettagliate sulle esecuzioni dei processi, che alla necessità di migliorare e supportare i processi di business in contesti sempre più competitivi e in rapida evoluzione. Obiettivo di questo articolo è quello di promuovere l'interesse per questa disciplina, introducendo un insieme di principi guida legati all'applicazione del process mining e individuando nuove importanti sfide per il futuro.

Keywords: Business Process Management, Tecniche di estrazione dei modelli di processo, Data Mining, Business Process Improvement, Total Quality Management

1. Introduzione

Il *process mining* costituisce un'area di ricerca relativamente giovane, che si posiziona da un lato tra la *computational intelligence* ed il *data mining* e dall'altro tra la modellazione e l'analisi dei *processi di business* (per una breve introduzione al concetto di processo e termini correlati, si veda il Riquadro 1). L'idea di base del process mining è quella di *modellare, monitorare e migliorare* i processi estraendo conoscenza dai log, oggi ampiamente disponibili nei sistemi informativi (Figura 1). I log contengono infatti informazioni legate all'esecuzione dei processi *nel mondo reale* che



sono di vitale importanza per la definizione di strategie per migliorare la qualità dei processi e ridurre i costi. Nel concreto, le applicazioni di process mining consentono: l'estrazione (automatica) di un modello di processo a partire da un log (*discovery*); la verifica di conformità (*conformance checking*), cioè l'individuazione di eventuali discrepanze tra un modello di processo e le informazioni contenute in un log; l'identificazione di reti sociali (*social network*) e organizzative; la costruzione automatica di modelli di simulazione; l'estensione e la revisione di modelli; la predizione delle possibili future evoluzioni di un'istanza di processo; l'estrazione (sulla base di dati storici) di raccomandazioni su come procedere nel corso di un'istanza di processo per raggiungere determinati obiettivi. Il process mining è quindi una tecnologia che supporta, di fatto, varie tecniche legate alla *Business Intelligence* (BI). Una di queste tecniche consiste, ad esempio, nel *Business Activity Monitoring* (BAM), che consente il monitoraggio in tempo reale di processi di business. Sempre nell'ambito della BI, il process mining può essere utilizzato a supporto del *Complex Event Processing* (CEP), che si riferisce all'analisi di grandi quantità di eventi e all'uso di questi per il monitoraggio, l'indirizzamento e l'ottimizzazione in tempo reale del business di un'organizzazione. Infine, il process mining può essere usato nell'ambito del *Corporate Performance Management* (CPM), ovvero per la misurazione delle performance di un processo o di un'organizzazione. Il process mining può essere utilizzato anche come piattaforma tecnologica sulla quale realizzare meccanismi di gestione dei processi quali il *Continuous Process Improvement* (CPI), il *Business Process Improvement* (BPI), il *Total Quality management* (TQM), e il *Six Sigma*. Tutte queste tecniche condividono l'idea che un processo vada "analizzato al microscopio" per identificare possibili miglioramenti, e il process mining nasce proprio con l'idea di fornire una tecnologia per supportare questa analisi.

Le aziende oggi stanno iniziando a porre molta enfasi anche sulla *corporate governance*, sulla *gestione del rischio*, e sulla *conformità*. Regolamentazioni, quali il Sarbanes-Oxley Act (SOX) e l'accordo Basilea II, per esempio, si focalizzano specificatamente sulla conformità della gestione aziendale a regole e norme condivise. Le tecniche di process mining offrono strumenti per rendere i controlli di conformità più rigorosi e per accertare la validità e l'affidabilità delle informazioni riguardanti i processi chiave di un'organizzazione.

Tutte queste problematiche possono beneficiare del process mining grazie anche al fatto che gli algoritmi sviluppati sono stati implementati in diversi sistemi, sia accademici che commerciali. Inoltre, ad oggi esiste un attivo gruppo di ricercatori che lavora sullo sviluppo di tecniche di process mining, che pertanto sta diventando uno degli "argomenti caldi" nella ricerca sul *Business Process Management* (BPM). Parallelamente, l'industria si sta dimostrando molto interessata e ricettiva rispetto a questi temi e sempre più produttori software stanno introducendo funzionalità di process mining nei loro prodotti. Esempi di tali prodotti sono: ARIS Process Performance Manager (Software AG), Comprehend (Open Connect), Discovery Analyst (StereoLOGIC), Flow (Fourspark), Futura Reflect (Futura Process

Intelligence), Interstage Automated Process Discovery (Fujitsu), OKT process mining suite (Exeura), Process Discovery Focus (Iontas/Verint), ProcessAnalyzer (QPR), ProM (TU/e), Rbminer/Dbminer (UPC), e Perceptive Reflect (Perceptive Software)

L'interesse per il process mining, sia nell'università che nel mondo dell'industria, è dimostrato anche dal fatto che *IEEE* ha creato una specifica Task Force con lo scopo di promuovere e supportare il process mining (si veda il Riquadro 2). Un importante risultato della task force è la redazione del *Manifesto sul Process Mining*, disponibile in diverse lingue, incluso l'italiano [2]. Il Manifesto è un documento che descrive le principali caratteristiche del process mining e fornisce una serie di linee guida per coloro che vogliono lavorare in questo campo nonché le più interessanti sfide per il futuro.

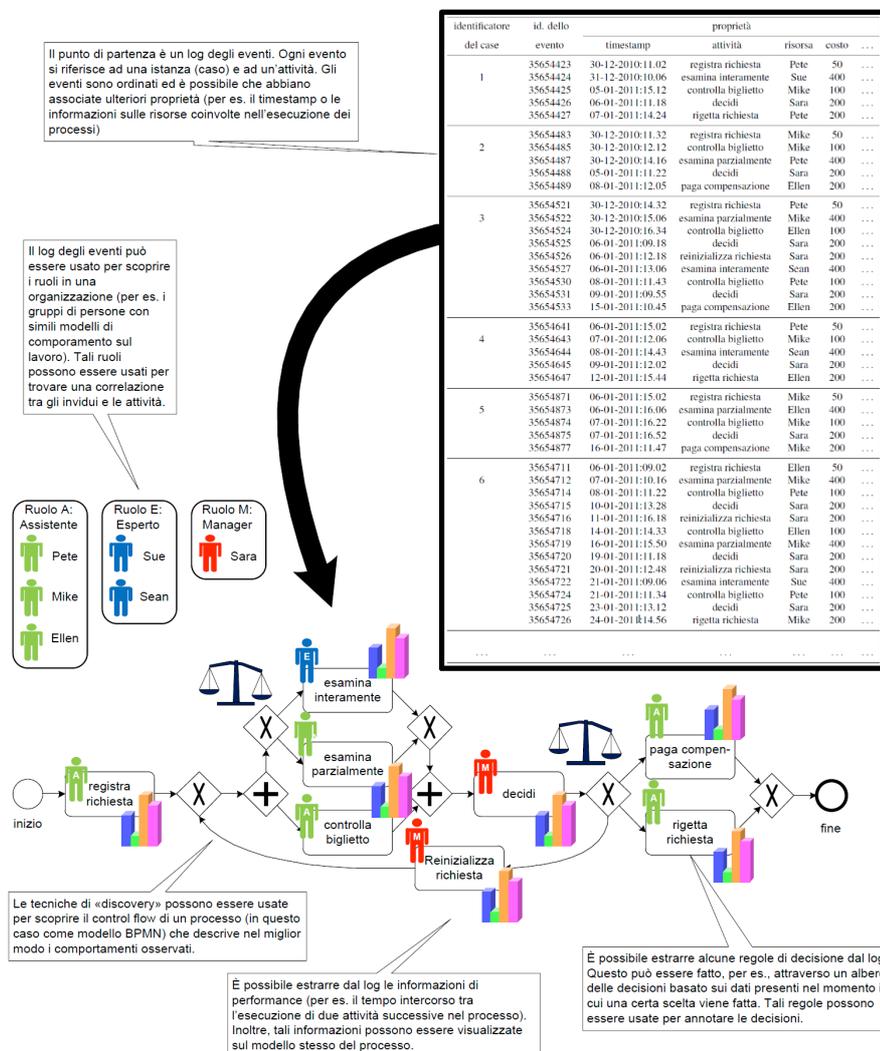


Figura 1
 Le tecniche di process mining estraggono conoscenza dai log al fine di modellare, monitorare e migliorare i processi di business.



2. Cos'è il *process mining*

La capacità di espansione dei sistemi informativi e di altri sistemi computazionali è caratterizzata dalla legge di Moore. Gordon Moore, il cofondatore di Intel, aveva previsto, nel 1965, che il numero di componenti in circuiti integrati sarebbe raddoppiato ogni anno. Durante gli ultimi 50 anni, questa crescita è stata effettivamente esponenziale sebbene leggermente più lenta. Questi progressi hanno condotto ad un'incredibile espansione dell'"universo digitale" (cioè di tutti i dati immagazzinati o scambiati elettronicamente) che, col passare del tempo, si sta via via allineando a quello reale.

La crescita di un universo digitale allineato con i processi di business rende possibile l'analisi di ciò che accade nella realtà sulla base di quanto registrato in un log. Un log può contenere eventi di vario tipo: un utente che ritira del denaro contante da uno sportello automatico, un dottore che regola un apparecchio per i raggi X, una persona che fa richiesta per una patente di guida, un contribuente che sottomette una dichiarazione dei redditi o un viaggiatore che riceve il numero di un biglietto elettronico sono tutti scenari in cui l'azione viene tracciata in un log. Si propone perciò la sfida di cercare di sfruttare questi dati in modo significativo, per esempio per fornire suggerimenti durante l'esecuzione di un processo, identificare colli di bottiglia, prevedere problemi nell'esecuzione, registrare violazioni, raccomandare contromisure e dare "forma" ai processi. Lo scopo del process mining è fare esattamente tutto questo.

Il punto di partenza per qualsiasi tecnica di process mining è sempre un *log degli eventi* (di seguito denominato semplicemente log). Tutte le tecniche di process mining assumono che sia possibile registrare eventi sequenzialmente in modo che ciascuno di questi si riferisca ad una determinata *attività* (cioè ad un passo ben definito di un processo) e sia associato ad una particolare istanza di processo.

Un'istanza di processo, o *case*, è una singola esecuzione del processo. Per esempio si consideri il processo di gestione di prestiti elargiti da un istituto di credito: ogni esecuzione del processo è intesa a gestire una richiesta di prestito.

I log possono contenere anche ulteriori informazioni circa gli eventi. Di fatto, quando possibile, le tecniche di process mining usano informazioni supplementari come le *risorse* (persone e dispositivi) che eseguono o che danno inizio ad un'attività, i *timestamp* o altri dati associati ad un evento (come la dimensione di un ordine).

2.1. Come si applicano le tecniche di process mining in pratica?

Come mostrato in Figura 2, il process mining prevede tre tipi di utilizzo il primo dei quali è detto *process discovery*: dato un log, le tecniche di discovery estraggono un modello di processo che è conforme con il comportamento registrato in tale log. Il secondo tipo è il *conformance checking*: un modello di processo che descrive il comportamento teorico atteso è confrontato con il comportamento reale del processo come registrato nel log per verificare se esistono delle divergenze. Il terzo tipo è

l'*enhancement* (miglioramento). In tal caso, l'idea è quella di estendere o migliorare un modello di processo esistente usando le informazioni contenute nei log. Mentre il conformance checking misura quanto un modello è allineato con ciò che accade nella realtà, questo terzo tipo di process mining si propone di cambiare o estendere il modello preesistente per adeguarlo alla realtà.

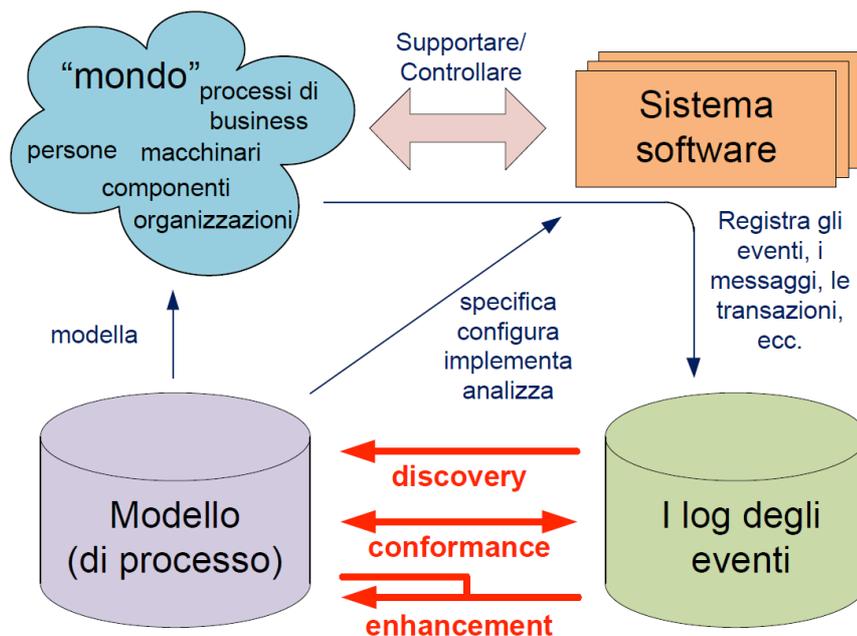


Figura 2

*I tre tipi principali di process mining:
(a) discovery, (b) conformance checking, e (c) enhancement.*

La Figura 3 descrive i tre tipi di process mining in termini di *input/output*. Le tecniche di discovery prendono in input un log e producono un modello. Il modello estratto è tipicamente un modello di processo (per esempio una rete di Petri, un modello BPMN, un modello EPC o un diagramma UML delle attività). Tuttavia, il modello può anche descrivere altre prospettive (come per esempio una social network che descrive la rete sociale di un'organizzazione). Le tecniche di conformance checking prendono in input un log e un modello. L'output consiste in una serie di informazioni diagnostiche che mostrano le differenze tra il modello e il log. Anche le tecniche di enhancement (revisione o estensione) richiedono un log e un modello in input. L'output è il modello stesso, migliorato o esteso.

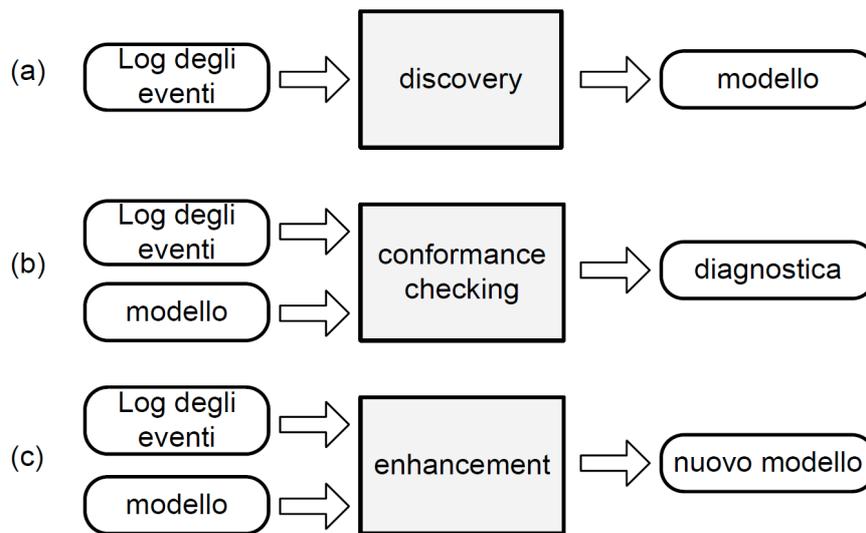


Figura 3

*I tre tipi base di process mining spiegati in termini di input e output:
(a) discovery, (b) conformance checking, e (c) enhancement.*

2.2. Che cosa non è il process mining

Esistono alcuni fraintendimenti comuni quando si parla di process mining. Alcuni produttori, analisti e ricercatori tendono a pensare alle tecniche di process mining come una forma specifica di data mining applicata *post mortem* sui log di istanze di processo già concluse. Tuttavia, questa visione non è corretta, in quanto:

1. *Il process mining non si limita alla scoperta del flusso di controllo.* Il discovery del flusso di controllo è spesso visto come la parte più interessante del process mining. Tuttavia, il process mining non si limita a questo. Il discovery è infatti solo una delle tre principali forme di process mining (assieme a conformance ed enhancement) ed, inoltre, non si limita unicamente al flusso di controllo: la prospettiva organizzativa, di istanza e temporale svolgono un ruolo altrettanto importante.
2. *Il process mining non è solo una particolare forma di data mining.* Il process mining può essere considerato come "l'anello mancante" fra il data mining ed il tradizionale model-driven BPM, in cui si parte da un modello di processo progettato manualmente. La maggior parte delle tecniche di data mining non sono infatti orientate ai processi. I modelli di processo hanno caratteristiche specifiche, come la possibilità di rappresentare comportamenti concorrenti, incomparabili con le strutture che tipicamente caratterizzano il data mining, quali alberi di decisione e regole associative. Questa discrepanza richiede lo sviluppo di modelli di rappresentazione e algoritmi completamente nuovi.

3. *Il process mining non si limita ad una analisi offline.* Le tecniche di process mining estraggono conoscenza a partire da dati storici. Anche se è possibile effettuare un'analisi a partire da dati "post mortem", il risultato di tale analisi può essere ottenuto anche considerando le istanze in esecuzione, allo scopo di fornire un supporto operativo alle decisioni (*operational support*). Ad esempio, un modello di processo precedentemente estratto da un log può essere successivamente impiegato per predire il tempo di completamento di un ordine iniziato da un cliente.

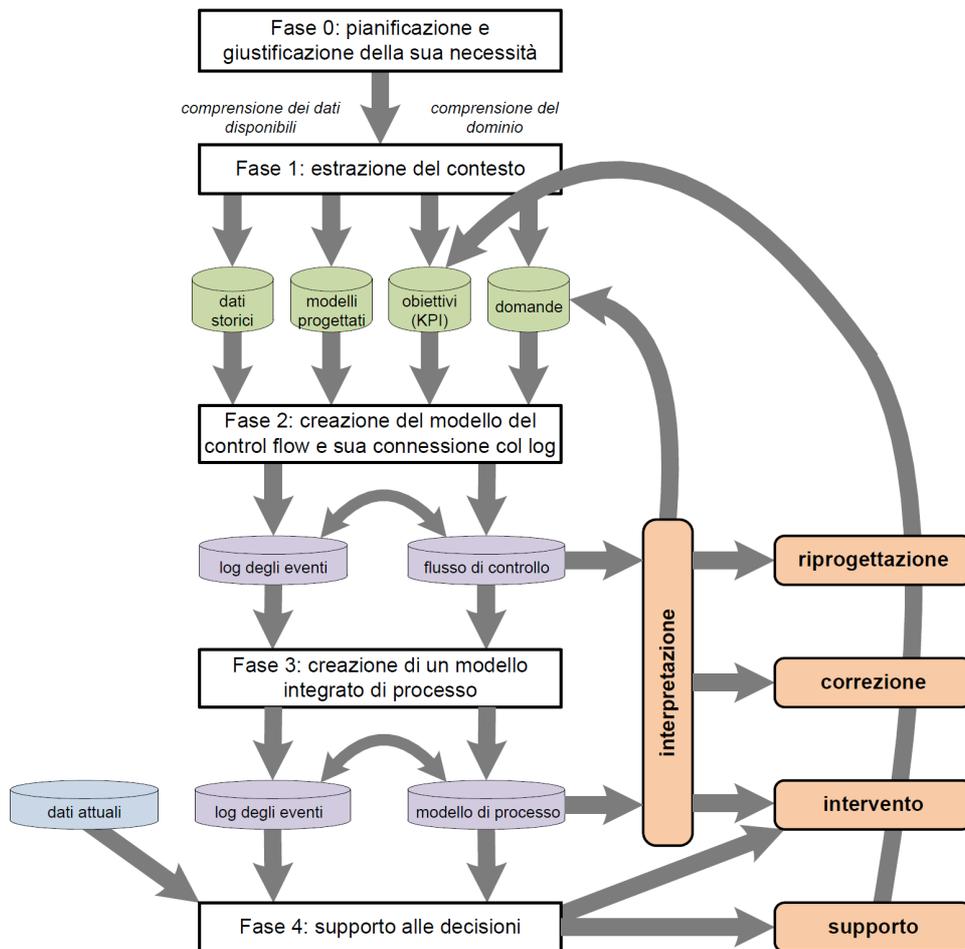


Figura 4

Ciclo di vita di un modello L che descrive un progetto di process mining costituito da cinque fasi: pianificazione e giustificazione (fase 0), estrazione (fase 1), creazione di un modello di flusso di controllo e connessione al log (fase 2), creazione di un modello di processo integrato (fase 3) e supporto alle decisioni (fase 4).*



2.3. Un framework di riferimento per l'applicazione del process mining

La Figura 4 riporta le principali fasi attraverso le quali si sviluppa un'analisi comprensiva dei processi di business basata sul process mining. Il primo passo consiste nella pianificazione e in uno studio atto a giustificare la pianificazione stessa (fase 0). Dopo l'avvio del progetto, è necessario interrogare sistemi informativi, esperti di dominio e manager per ricavare dati, modelli, obiettivi e quesiti ai quali si vuole dare risposta con l'analisi (fase 1). Questa attività richiede una comprensione dei dati che si hanno a disposizione ("quali dati si possono usare per l'analisi?") e del dominio ("quali sono i quesiti rilevanti?"), e fornisce i risultati riportati in Figura 4 (cioè dati storici, modelli progettati, obiettivi, quesiti). Durante la fase 2 viene (ri)costruito il modello del flusso di controllo e lo si collega al log. In questa fase si possono usare tecniche automatiche di discovery. Il modello estratto può già essere utilizzato per rispondere ad alcuni dei quesiti posti, e di conseguenza avviare una fase di adattamento e rimodellazione del processo. Inoltre, il log può essere filtrato o adattato sulla base del modello (ad esempio rimuovendo attività rare o istanze anomale, ed inserendo eventi mancanti). Quando il processo è piuttosto strutturato, il modello del flusso di controllo può essere esteso con altre prospettive durante la fase 3. Per esempio, è possibile identificare le relazioni tra le risorse che concorrono all'esecuzione di un processo (*prospettiva dell'organizzazione*), la frequenza con cui certi eventi accadono e quali attività costituiscono colli di bottiglia (*prospettiva del tempo*).

La relazione tra log e modello, stabilita durante la fase 2, può essere impiegata per estendere il modello stesso (ad esempio, i timestamp possono essere utilizzati per stimare i tempi di attesa delle attività).

Infine, i modelli costruiti durante la fase 3 si possono utilizzare per fornire supporto alle decisioni (fase 4), combinando la conoscenza estratta a partire dai dati storici con informazioni sulle istanze in esecuzione. In questo modo è possibile generare raccomandazioni su cosa fare, generare informazioni predittive sul futuro andamento del processo, e intervenire sul processo stesso. È opportuno evidenziare che le fasi 3 e 4 possono essere eseguite solo se il processo è sufficientemente stabile e strutturato.

2.4. Applicazioni pratiche e industriali

Negli ultimi dieci anni, le tecniche di process mining sono state utilizzate in più di 100 organizzazioni inclusi comuni (per es., Alkmaar, Zwolle, Heusden in Olanda), agenzie governative (per es., Rijkswaterstaat, Centraal Justitieel Incasso Bureau), banche (per es., ING Bank), ospedali (per es., Catharina hospital ad Eindhoven), multinazionali (per es., Deloitte), industrie manifatturiere e loro clienti (per es., Philips, ASML, Ricoh, Thales). Questa diffusione dimostra l'ampio numero di contesti nei quali è possibile applicare il process mining.

Recentemente, l'università di Eindhoven ha iniziato a collaborare al progetto CoSeLog centrato sul process mining e che coinvolge 10 comuni olandesi (cfr. <http://www.win.tue.nl/coselog>). In questo progetto, le tecniche di process mining vengono utilizzate per scoprire similarità e differenze tra i



comuni nella gestione dei loro compiti. L'obiettivo ultimo è quello di creare un'infrastruttura centralizzata e condivisa che garantisca la comunicazione tra i comuni e che realizzi, per quanto è possibile, la standardizzazione dei processi che i comuni stessi eseguono. I primi risultati sono estremamente incoraggianti: per maggiori informazioni, si rimanda a [1] e a [4]. Un altro interessante caso di studio ha coinvolto il dipartimento nazionale per i lavori pubblici in Olanda [3]. Qui, le tecniche di process mining sono state utilizzate per generare un modello di processo. Il log utilizzato per questo caso di studio contiene 14280 istanze di processo e complessivamente 150000 eventi; ciò dimostra che molte tecniche di process mining scalano bene su log reali. Nell'ambito dello stesso caso di studio, a partire dalle informazioni contenute nel log, è stato possibile costruire una rete sociale che ha enfatizzato come la responsabilità dell'esecuzione delle istanze di processo passasse da un attore ad un altro. Gli esempi appena descritti si riferiscono a processi che sono ben strutturati, spesso noti in letteratura come "processi a lasagna". Tuttavia, altri processi sono caratterizzati da una struttura molto più complessa (i cosiddetti "processi a spaghetti"). Anche se non tutte le tecniche di process mining esistenti possono essere applicate a questo tipo di processi, esistono alcune tecniche che sono state sviluppate in maniera specifica per l'analisi dei processi a spaghetti. Queste risultano estremamente utili, in quanto permettono di migliorare processi strutturalmente complessi, sia in termini di esecuzione che in termini di comprensibilità. Processi non strutturati si riscontrano spesso in domini altamente dinamici dove è richiesto un alto grado di flessibilità e in cui ogni esecuzione è differente dalle altre. Questo tipo di comportamento è spesso osservato nei processi per la fornitura di servizi sanitari e nel trattamento di pazienti ospedalizzati. L'articolo [7] riporta l'applicazione di tecniche di process mining nel reparto di oncologia di uno dei principali ospedali in Olanda: il log degli eventi utilizzato per questo caso di studio contiene 24331 eventi che si riferiscono a 376 diverse attività.

Un altro interessante caso studio [6] riguarda l'applicazione di tecniche di process mining combinate con tecniche di classificazione a log provenienti da un sistema di gestione di container in un porto italiano di Transshipment: il log utilizzato è dell'ordine di 50Mb di dati riguardanti il transito di 5336 container nei primi due mesi del 2006 e le relative movimentazioni su piazzale. La tecnica ha permesso di scoprire diversi scenari di gestione dei container (ognuno descritto con un modello dettagliato di processo) e, altresì, di evidenziare le correlazioni tra tali casi d'uso e alcune proprietà non strutturali dei processi stessi (es. porto di provenienza/destinazione, compagnia di navigazione, dimensione del container).

3. Principi guida e sfide per il futuro del *process mining*

Attualmente esistono tecniche e strumenti che permettono di realizzare tutte le fasi riportate in Figura 4. Tuttavia il process mining è un paradigma relativamente recente e la maggior parte degli strumenti disponibili non sono ancora maturi. Inoltre, i nuovi utenti spesso non hanno una conoscenza approfondita del potenziale e dei limiti del process mining. Per tali ragioni, vengono di seguito forniti alcuni principi guida ed alcune sfide per potenziali

utenti, ricercatori e sviluppatori interessati a far avanzare lo stato dell'arte in questa disciplina.

3.1. Principi guida

Come per ogni nuova tecnologia, quando il process mining viene applicato in scenari reali, è possibile incorrere in errori ricorrenti. La Tabella 1 riprende i sei principi guida del manifesto del process mining [2], che danno una serie di indicazioni da seguire per una corretta applicazione di queste tecniche.

Per esempio, si consideri il principio guida *PG4*: “Gli eventi devono essere legati ad elementi del modello”. Il conformance checking e l'enhancement si basano significativamente su tale relazione. Il conformance checking, per esempio, si basa sul principio del “replay”. Dopo aver collegato le attività nel modello con gli eventi nel log, questi ultimi sono ordinati per ordine di occorrenza (per es., usando i timestamp). Si può fare così il replay del log per verificare che esso sia conforme al modello.

PG1	Gli eventi devono essere trattati come entità di prima classe Gli eventi dovrebbero essere considerati attendibili nel senso che occorre assumere che gli eventi registrati nei log siano realmente avvenuti e che gli attributi associati agli eventi siano corretti. I log dovrebbero essere completi, cioè contenere tutti gli eventi rilevanti per descrivere un processo e ogni evento dovrebbe avere una precisa semantica. I dati associati agli eventi dovrebbero essere resi disponibili, tenendo in considerazione problematiche di privacy e sicurezza.
PG2	L'estrazione dei log deve essere guidata da quesiti Senza quesiti concreti, è difficile combinare eventi realmente significativi in un log. Si considerino, per esempio, le centinaia di tabelle che possono costituire il database di un sistema ERP (Enterprise Resource Planning) come SAP. Senza quesiti concreti, non è possibile stabilire come procedere nell'estrazione di un log.
PG3	Occorre supportare concorrenza, punti di decisione e altri costrutti di base legati al flusso di controllo Le tecniche di process mining dovrebbero supportare i “costrutti tipici”; questi annoverano, per esempio, attività sequenziali, parallele o ripetute, o anche la scelta di una data attività in un insieme di alternative sulla base di precondizioni.
PG4	Gli eventi devono essere collegati ad elementi del modello di processo Le tecniche di conformance checking ed enhancement si basano significativamente sulle relazioni tra gli elementi del modello di processo e gli eventi nei log. Tali relazioni possono essere usate per effettuare il “replay” degli eventi sul modello. Le tecniche di replay possono essere usate per rilevare discrepanze tra i log e i modelli di processo. Tali tecniche possono essere anche utilizzate per arricchire il modello con informazioni aggiuntive estratte dal log (per es., l'analisi dei timestamp nei log possono identificare colli di bottiglia durante l'esecuzione).
PG5	I modelli devono essere trattati come astrazioni utili della realtà che mettono in risalto alcuni aspetti Un modello derivato dai dati associati agli eventi fornisce una certa vista della realtà, che dovrebbe essere considerata come un'astrazione del comportamento catturato dal log. Tale astrazione non è, in generale, assoluta ma può essere utile avere diverse astrazioni in funzione del particolare punto di vista di interesse.
PG6	Il process mining deve essere un processo continuo Data la natura dinamica dei processi, non bisogna considerare il process mining come un'attività da eseguirsi una volta per tutte. L'obiettivo non dovrebbe, infatti, essere quello di costruire un modello “definitivo”. Al contrario, utenti e analisti dovrebbero ripetere le analisi con una certa periodicità.

Tabella 1
I sei principi guida descritti nel manifesto.

3.2. Le sfide del futuro

Il process mining è un'importante tecnologia per la gestione dei processi nelle moderne organizzazioni. Nonostante esista attualmente una vasta gamma di tecniche di process mining, ci sono ancora molte sfide ancora aperte.

La Tabella 2 elenca le undici sfide descritte nel manifesto [2]. Per esempio, nella sfida S4: "Gestire il *concept drift*", il termine *concept drift* si riferisce alla situazione in cui il processo cambia la propria struttura mentre viene analizzato. Per esempio, nella parte iniziale di un log un processo può prevedere che due attività siano concorrenti mentre nel seguito il processo cambia e la modellazione delle attività impone che queste vengano eseguite sequenzialmente. I processi possono cambiare per varie ragioni e i loro cambiamenti possono essere periodici e relativi ad un certo lasso temporale (per es., "in dicembre c'è più domanda" o "venerdì ci sono meno impiegati disponibili") oppure questi possono essere dovuti a condizioni ambientali che cambiano (per es., "il mercato è diventato più competitivo"). Tali cambiamenti hanno impatto sui processi, inducendo modifiche temporanee o permanenti. È quindi estremamente importante rilevare questi cambiamenti ed analizzarli. Purtroppo, la maggior parte delle tecniche esistenti parte dall'assunzione che la struttura dei processi non sia soggetta a modifiche nel tempo.

A causa dell'enorme crescita delle dimensioni dei log registrati dai sistemi software, sta diventando sempre più complessa l'analisi di tali log e l'estrazione di informazioni da questi. Nonostante gli ottimi risultati ottenuti dalle odierne tecniche per l'analisi automatica e il process mining, la complessità di tale analisi rende necessario includere l'intervento umano per interpretare e raffinare i risultati. In relazione alle sfide S8 and S9, le tecniche automatiche devono essere integrate con interfacce utente user-friendly ed intuitive e con metodi per l'analisi visuale dei risultati. In questo modo, gli analisti possono integrare la loro flessibilità, creatività e conoscenza del dominio con le tecniche automatiche di process mining in modo da arrivare ad ottenere una completa comprensione dei processi eseguiti in una data organizzazione. Il lavoro [5] discute alcuni approcci per integrare tecniche puramente automatiche e metodologie di analisi visuale.

S1	Costruzione, fusione e filtraggio dei log Quando vengono estratti gli eventi da un log per il loro uso come input di tecniche di process mining, ci sono svariate problematiche da tenere in considerazione: i dati possono essere distribuiti su una moltitudine di sorgenti eterogenee, questi possono essere incompleti, il log può contenere eventi che sono outliers o hanno diversi livelli di granularità, ecc.
S2	Manipolazione di log complessi e con caratteristiche diverse I log possono avere molte caratteristiche differenti. Alcuni log possono essere estremamente grandi, il che li rende difficile da gestire; altri possono essere così ridotti da non contenere dati sufficienti per giungere a conclusioni affidabili.
S3	Creazione di benchmark rappresentativi Sono necessari dei benchmark contenenti insiemi di dati significativi di alta qualità. Questo è necessario per essere in grado di comparare e migliorare algoritmi e applicazioni.

S4	Gestire il concept drift Un processo può modificarsi mentre viene analizzato. La comprensione del concept drift è di primaria importanza per la gestione e l'analisi dei processi.
S5	Migliorare i limiti di rappresentazione nel process discovery Un'attenta selezione del bias di rappresentazione è necessaria, al fine di garantire risultati di elevata qualità.
S6	Valutare i criteri di qualità Per valutare la qualità dei risultati di una tecnica di process mining, ci sono quattro dimensioni che spesso si escludono a vicenda: (a) <i>fitness</i> , (b) <i>simplicity</i> , (c) <i>precision</i> , e (d) <i>generalization</i> . Una chiara sfida è quella di scoprire modelli la cui qualità sia bilanciata rispetto a tutte le dimensioni.
S7	Cross-organizational mining Ci sono vari casi d'uso dove i log di molte organizzazioni sono disponibili per l'analisi. Alcune organizzazioni infatti cooperano per gestire i loro processi (per es., i partner nelle <i>supply chain</i>) oppure queste eseguono essenzialmente lo stesso processo condividendo esperienza, conoscenza o un'infrastruttura comune. Tradizionalmente le tecniche di process mining considerano ogni log come generato all'interno di una singola organizzazione. Una nuova sfida è quella di scoprire le relazioni e le interconnessioni tra i diversi partner che cooperano nell'esecuzione di dati processi.
S8	Fornire supporto alle decisioni L'applicazione di tecniche di process mining non si limita all'analisi off-line ma può essere usato per un supporto in real-time, ad esempio, per ottimizzare l'esecuzione dei processi. E' possibile identificare tre tipi di supporto alle decisioni: monitoraggio, predizione, e raccomandazione.
S9	Combinare il process mining con altri tipi di analisi La sfida, in questo caso, è quella di combinare le tecniche di process mining automatico con altri tipi di analisi (tecniche di ottimizzazione, data mining, simulazione, analisi visuale, etc.) al fine di estrarre ulteriori informazioni dai dati.
S10	Migliorare l'usabilità per gli utenti non esperti La sfida è quella di nascondere i sofisticati algoritmi di process mining dietro interfacce <i>user-friendly</i> che si auto-configurino e suggeriscano caso per caso quali sono le analisi più opportune.
S11	Migliorare la comprensibilità per gli utenti non esperti I risultati ottenuti da tecniche di process mining possono essere difficili da interpretare e, quindi, potrebbero portare a conclusioni errate. Al fine di evitare tali problemi, i risultati dovrebbero essere presentati usando una rappresentazione semplice, associandovi un livello di confidenza da cui gli stessi risultati sono caratterizzati.

Tabella 2

Alcune delle più importanti sfide nell'ambito del process mining.

3.3 Che cosa rende il *process discovery* una sfida?

Sebbene il process mining non riguardi solamente l'estrazione di un modello di processo a partire da un log, questo aspetto è sicuramente il più complesso ed interessante. In quest'ambito, data mining e process mining si scontrano con una serie di problematiche comuni, quali ad esempio la gestione del rumore, del concept drift e di grandi e complessi spazi di ricerca.

Tuttavia, il process discovery si trova ad affrontare una gamma di ulteriori sfide:

- non si hanno a disposizione esempi negativi (cioè un log mostra quello che è accaduto ma non quello che non dovrebbe mai accadere);
- a causa della concorrenza, della presenza di cicli e di scelte mutuamente esclusive, lo “spazio degli stati” possibili ha una dimensione molto grande ed il log contiene tipicamente solo una frazione molto piccola di tutti i comportamenti possibili;
- non c'è nessun nesso diretto tra la dimensione del modello e il suo comportamento: un modello può descrivere lo stesso numero di comportamenti di un modello strutturalmente più complesso;
- è necessario trovare il giusto bilanciamento tra i seguenti quattro criteri qualitativi, spesso in contrasto tra loro (si veda la sfida S6):
 - a) *fitness* (capacità di riprodurre i comportamenti osservati),
 - b) *simplicity* (la riduzione di grandi e complessi modelli a strutture più semplici),
 - c) *precision* (per evitare “*underfitting*”), e
 - d) *generalization* (per evitare “*overfitting*”).¹

Per illustrare concretamente alcune di queste sfide, si consideri il modello di processo in Figura 5. Il modello consiste in una rete di Petri che descrive un processo che inizia con l'attività **a** e finisce con **d**. Nel mezzo, **k** attività possono essere eseguite in parallelo. Lo *i*-esimo ramo parallelo contiene un punto di scelta sull'esecuzione di **b_i**, oppure **c_i**. Ne consegue che il modello esibisce **2^k k!** comportamenti distinti; con **k=10**, questo significa che il processo supporta **3.715.891.200** sequenze di esecuzione possibili. Per esempio, sono ammissibili sia **a c₅ b₃ c₁ b₂ b₄ c₆ c₈ b₇ c₉ c₁₀ d** che **a b₁ c₂ b₃ c₄ b₅ c₆ b₇ c₈ b₉ c₁₀ d**. Costrutti di concorrenza e di scelta mutuamente esclusiva generalmente causano un'esplosione del numero di tracce ammissibili. L'introduzione di cicli di attività (ripetizioni) nel modello può addirittura portare il numero di tracce ammissibili a diventare infinito. Non è di conseguenza realistico assumere che tutte le tracce supportate dal processo vengano effettivamente osservate in un log.

Fortunatamente, gli algoritmi esistenti di *process discovery* non richiedono l'osservazione completa di tutte le possibili combinazioni di sequenze per scoprire attività concorrenti. Per esempio, il classico algoritmo α [1] è in grado di derivare una rete di Petri sulla base di meno di **4k(k-1)** tracce. L'algoritmo α ha unicamente bisogno di avere a disposizione tutte “le successioni di attività” piuttosto che tutti i modi in cui le attività possono essere intervallate; in altre parole, l'algoritmo richiede che nel caso in cui l'attività **x** possa essere seguita direttamente da **y**, tale comportamento sia presente nel log almeno una volta. Viceversa, le tecniche di “*knowledge discovery*” tradizionali non riescono a scoprire modelli di processo come

¹ In letteratura, un modello estratto con tecniche di process mining è *overfitting* se non generalizza e permette tutti e soli i comportamenti riscontrati nel log. Viceversa, un modello è *underfitting* se è troppo generale e considera certi comportamenti come ammissibili sebbene non ci sia nessuna evidenza che questi debbano essere supportati.

quello di Figura 5. Si ritiene, infine, che il recente trend delle tecniche di *process improvement* (Six Sigma, TQM, CPI, CPM, etc.) e di *compliance* (SOX, BAM, etc.) possano trarre enormi benefici dall'utilizzo del process discovery. Per questa ragione, la speranza è quella che questo articolo stimoli la comunità italiana di ICT a sviluppare nuove tecniche di "knowledge discovery" che pongano i processi di business al centro dell'analisi. Le sfide per il futuro descritte in questo articolo sono dettagliate nel manifesto sul process mining [2]. Un'estesa analisi dello stato dell'arte sull'argomento è anche disponibile in [1].

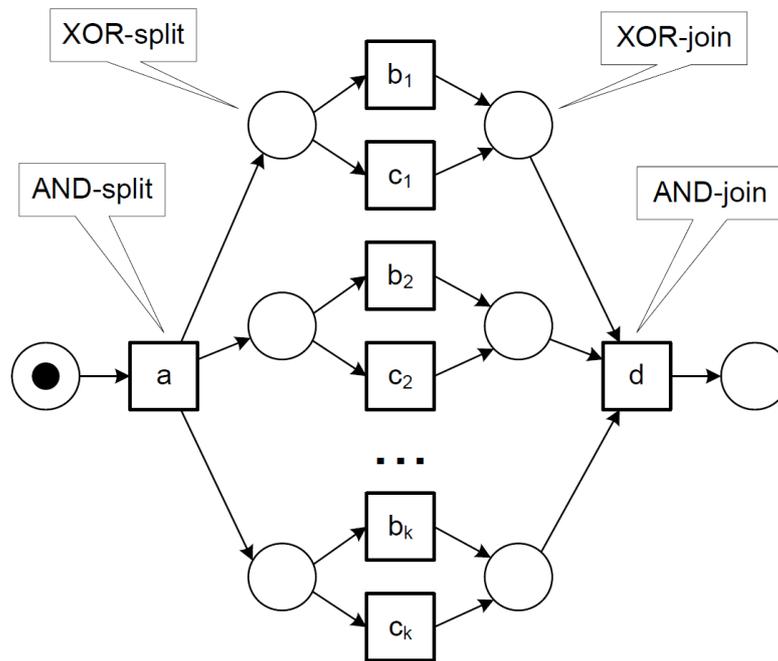


Figura 5
Una rete di Petri con $2^k k!$ diverse sequenze di esecuzione possibili.



Riquadro 1 - Introduzione ai principali concetti legati ai processi di business e alla loro gestione

Oggi, le organizzazioni cercano di rendere i loro processi interni sempre più efficienti così da minimizzare i costi e i tempi per la loro esecuzione per fronteggiare un mercato che diventa sempre più globalizzato e competitivo. Per questa ragione i sistemi software di gestione di processi (come SAP o ERP) stanno diventando sempre più richiesti perchè riescono ad ottimizzare e controllare l'esecuzione dei processi. In [8], Weske definisce un processo di business come una serie di attività eseguite in maniera coordinata all'interno di un contesto organizzativo e/o tecnico con lo scopo di perseguire un certo obiettivo aziendale. È necessario esprimere il *modello* di processo esplicitamente per poterne orchestrare concretamente l'esecuzione in sistemi software. In particolare, occorre descrivere quali vincoli ne disciplinano l'esecuzione, quali dati vengono consumati e prodotti dalle attività, nonché eventuali vincoli organizzativi che disciplinano chi può/deve eseguire cosa. Uno stesso modello di processo viene tipicamente messo in esecuzione più volte ed ogni esecuzione viene chiamata *istanza* o *case*. Per es., una banca esegue una nuova istanza di un processo per elargire un prestito ogni volta che un nuovo cliente effettua una richiesta.

Tipicamente un log contiene una serie di *tracce di esecuzione* tutte diverse tra di loro, ciascuna delle quali può essere replicata una o più volte in istanze di processo.

Riquadro 2 - IEEE Task Force sul *process mining*

Nel 2009 è stata fondata la *task force sul process mining* sotto il patrocinio dell'*Institute of Electrical and Electronic Engineers, Inc.* (IEEE). Scopo principale della task force è promuovere l'uso delle tecniche e degli strumenti di process mining e stimolarne nuove applicazioni. La task force si prefigge a tutt'oggi una serie di obiettivi specifici, fra i quali:

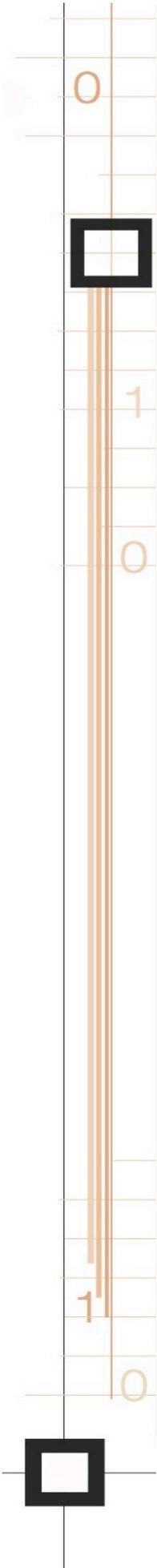
- diffondere lo stato dell'arte sul process mining a utenti, sviluppatori, consulenti, manager e ricercatori;
- partecipare attivamente alla standardizzazione di come gli eventi debbano essere rappresentati in un log;
- organizzare *tutorial*, *special session*, *workshop* e *panel* nell'ambito di conferenze e convegni internazionali;
- pubblicare articoli, libri, video ed edizioni speciali di riviste sul tema.

La Task Force comprende più di 10 produttori di software, tra i quali si trovano HP, Fujitsu e IBM, oltre 20 istituti di ricerca, incluse le università di affiliazione degli autori di questo articolo, molte aziende di consulenza (come Deloitte) e semplici utenti (per es., Rabobank).

Fin dalla sua fondazione, la Task Force ha svolto varie attività riconducibili agli obiettivi sopradescritti, tra cui la pubblicazione del primo libro sul process mining [1], un sito web (www.processmining.org), molti workshop, sessioni a conferenze e scuole di approfondimento. Nel 2010, la Task Force ha standardizzato XES (www.xes-standard.org), un formato per la memorizzazione dei log, estendibile e supportato dalla *libreria OpenXES* (www.openxes.org) e da strumenti di process mining quali ProM, XESame, Nitro, etc. All'indirizzo <http://www.win.tue.nl/ieeetfpm/> è possibile reperire maggiori informazioni sulle attività della Task Force.

Bibliografia

1. W.M.P. van der Aalst. *Process Mining: Discovery, Conformance and Enhancement of Business Processes*. Springer-Verlag, Berlin, 2011.
2. AA.VV. *Manifesto del Process Mining*. IEEE Task Force on Process Mining.
<http://www.win.tue.nl/ieeetfpm/lib/exe/fetch.php?media=shared:pmm-italian-v2.pdf>
3. W.M.P. van der Aalst, H.A. Reijers, A.J.M.M. Weijters, B.F. van Dongen, A.K. Alves de Medeiros, M. Song, H.M.W. Verbeek *Business process mining: An industrial application* Journal Information Systems, Volume 32 Issue 5, July, 2007, Pages 713-732, Elsevier Science Ltd. Oxford, UK
4. J. C. A. M. Buijs, and B. F. van Dongen and W. M. P. van der Aalst. *Towards Cross-Organizational Process Mining in Collections of Process Models and Their Executions* Business Process Management Workshops, Lecture Notes in Business Information Processing, 2012, Volume 100, Part 1, 2-13

- 
- 
5. M. de Leoni, W.M.P. van der Aalst, A.H.M. ter Hofstede. *Process Mining and Visual Analytics: Breathing Life into Business Process Models*. Alexandru Floares (Ed.), Computational Intelligence. Nova Science Publishers, Hauppauge, USA, 2012 (In preparazione)
 6. F. Folino, G. Greco, A. Guzzo, L. Pontieri. *Mining usage scenarios in business processes: Outlier-aware discovery and run-time prediction*. Data Knowledge Engineering, Volume 70, nr 12, Pages 1005-1029, 2011. Elsevier.
 7. R. S. Mans, M. H. Schonenberg, M. Song, W. M. P. van der Aalst and P. J. M. Bakker. *Application of Process Mining in Healthcare A Case Study in a Dutch Hospital* Communications in Computer and Information Science, 1, Volume 25, Biomedical Engineering Systems and Technologies, Part 4, Pages 425-438 Springer-Verlag, Berlin, 2009.
 8. M.Weske. *Business Process Management: Concepts, Languages, Architectures*. Springer-Verlag, Berlin, 2007.

Biografie

Wil van der Aalst è professore ordinario di Information System presso la Technische Universiteit Eindhoven (TU/e). Attualmente è anche un “adjunct professor” alla Queensland University of Technology (QUT), dove lavora presso il gruppo di Business Process Management (BPM). I suoi interessi di ricerca vertono su process mining, Petri net, business process management, modellazione e analisi di processi di business. Ha un H-index di 90 secondo Google Scholar, così da essere il ricercatore olandese in informatica con il più alto H-index. Le sue idee hanno influenzato ricercatori, sviluppatori software e comitati di standardizzazione che lavorano sul supporto ai processi di business.

W.M.P.v.d.Aalst@tm.tue.nl

Andrea Burattin ha conseguito la Laurea Magistrale in Informatica nel 2009, presso l'Università degli Studi di Padova e sta completando il Dottorato di Ricerca presso la Scuola di Dottorato in Scienze e Ingegneria dell'Informazione consorziato fra le Università di Padova e Bologna. Ha trascorso alcuni periodi (nov. 2010 e da ott. 2011 fino a gen. 2012) presso la Technische Universiteit Eindhoven (TU/e). La sua attività di ricerca riguarda lo studio delle problematiche legate al deploy di strumenti di process mining, configurazione automatica di parametri per algoritmi di mining, control-flow discovery e stream process mining.

burattin@math.unipd.it

Antonella Guzzo è Ricercatore confermato presso la Facoltà di Ingegneria dell'Università della Calabria ed ha conseguito il Dottorato di Ricerca in Ingegneria dei Sistemi ed Informatica, presso il Dip. DEIS dell'UNICAL. Ha ricoperto numerosi incarichi di collaborazione scientifica con l'Istituto di Calcolo e Reti ad Alte Prestazioni (ICAR-CNR) di cui è stata associata di ricerca.



La sua attività di ricerca riguarda le tecniche di Process Mining, Data Mining e Knowledge Representation. Su questi temi, è (co)autore di più di 40 pubblicazioni scientifiche tra articoli su riviste e conferenze internazionali e nazionali con referee.

guzzo@deis.unical.it

Massimiliano de Leoni ha conseguito un dottorato di ricerca in Ingegneria Informatica presso SAPIENZA – Università di Roma. Attualmente, è titolare di un assegno di post-dottorato presso la Technische Universiteit Eindhoven (TU/e). È stato visitatore scientifico presso la Queensland University of Technology (QUT) e anche presso la RMIT University, Melbourne. I suoi interessi di ricerca si focalizzano sul workflow management e process mining, con particolare attenzione su tecniche di “business process visualization” e di process mining centrate sui dati e le risorse, nonché su framework per l’adattabilità automatica di processi di business.

m.d.leoni@tue.nl

Fabrizio Maggi ha conseguito un dottorato di ricerca in Informatica nel 2010 presso l’università di Bari. Attualmente, è titolare di un assegno di post-dottorato presso la Technische Universiteit Eindhoven (TU/e) nel gruppo di ricerca Architecture of Information Systems (AIS). La sua tesi di dottorato è intitolata "Modellazione, Implementazione e Miglioramento dei Processi di Business". È autore di più di 20 pubblicazioni riguardanti process mining, revisione di modelli di processo, modelli di processo dichiarativi, regole di business, monitoraggio di business constraints, service oriented architectures, coreografie e composizione di servizi.

f.m.maggi@tue.nl

Marco Montali è Ricercatore a Tempo Determinato presso la Libera Università di Bozen-Bolzano. Nel 2009, ha conseguito il Dottorato di Ricerca in Ingegneria Elettronica, Informatica e delle Telecomunicazioni, presso l’Università di Bologna. La sua tesi di Dottorato è stata insignita del premio "Marco Cadoli" come miglior tesi discussa tra il 2007 e il 2009 su tematiche di logica computazionale. È autore di più di 50 pubblicazioni su aspetti teorici, metodologici e sperimentali legati alla specifica formale, la verifica e il monitoraggio di processi di business centrati sui dati, sistemi multi-agente, linee guida cliniche e sistemi orientati ai servizi.

montali@inf.unibz.it