

# Sistemi Informativi (IS) come Disciplina

**Obiettivo:** La disciplina dei Sistemi Informativi (IS) si occupa di studiare, categorizzare e valutare metodologie per lo sviluppo di sistemi informativi all'interno delle organizzazioni.

**Origini e Evoluzione:** Nata inizialmente nelle scuole di economia aziendale, si è poi estesa ad altre aree, come l'informatica (Computer Science). Ad esempio, il process mining è un approccio nato dall'informatica per unire la Data Science e il Business Process Management.

## Comunità di Riferimento:

- **Comunità scientifica:** Include numerose conferenze internazionali (ICIS, ECIS, BPM, ICMP), riviste scientifiche e altre pubblicazioni.
- **Comunità industriale:** Comprende metodologie e strumenti pratici come i sistemi ERP, BPM, PLM e HRM.

## Obiettivi dei Sistemi Informativi

1. **Allineamento IT-Business:** Sviluppare metodologie per allineare le scelte di progettazione IT con gli obiettivi aziendali. Questo è considerato il fine principale della disciplina. Esempi includono il Business Process Management, l'Enterprise Architecture e i modelli di allineamento strategico.
2. **Design Science:** È un approccio costruttivo che si concentra sulla creazione di artefatti (es. modelli, sistemi, metodi) per migliorare le prestazioni. Mappa lo spazio degli obiettivi a quello degli attributi. Se un artefatto non funziona, si possono modificare l'artefatto o i requisiti. Non è una scienza esplicativa, ma progettuale.
3. **Management Science:** Studia strumenti e metodi per supportare il processo decisionale nelle organizzazioni. Include tecniche analitiche, modelli matematici e indicatori di performance (KPI). È interdisciplinare e si collega a management, economia, statistica, ricerca operativa.
4. **Information Management:** Si occupa del ciclo di vita delle informazioni: acquisizione, archiviazione, gestione, protezione, distribuzione. È alla base della Business Intelligence, che si basa su dati empirici per supportare le decisioni.
5. **IT Trends:** Riguarda le tecnologie emergenti che influenzano l'evoluzione dei Sistemi Informativi. Include Big Data, AI, Process Mining, BI 2.0, sistemi ERP, BPM, PLM, HRM. I trend aiutano a migliorare i processi aziendali e il supporto decisionale.
6. **CS Methods (metodi dell'informatica):** Comprende le basi dell'informatica utili alla progettazione dei SI: strutture dati, logica, statistica, programmazione. Si applica anche il metodo scientifico ODME: Osservazione, Design, Misurazione, Valutazione.

## Concetti Chiave: Il Metodo ODME

La nozione centrale nella disciplina IS è l'utilizzo del metodo scientifico nella progettazione di Sistemi Informativi a supporto delle attività aziendali e organizzative. Questo approccio si articola nelle seguenti fasi, spesso indicate con l'acronimo ODME:

- **Observe (Osservare):** Questa fase implica l'osservazione della realtà aziendale e dei processi esistenti, identificando problemi, esigenze e definendo gli obiettivi. Tecniche di

osservazione possono includere l'analisi di indicatori chiave di prestazione (KPI) e log degli eventi. Ogni osservazione relativa all'elaborazione delle informazioni deve essere coerente con le specifiche e gli obiettivi.

- In Business Intelligence (BI) corrisponde alla fase di business and data understanding.
- In Process Mining, l'osservazione si basa sui log degli eventi che riflettono il comportamento reale dei processi.
- **Design (Progettare):** Durante questa fase si definiscono le specifiche del sistema informativo e si progettano gli artefatti (modelli, processi, interfacce) che dovrebbero soddisfare gli obiettivi identificati. Le specifiche sono essenziali per documentare il ciclo di vita dello sviluppo e per fornire input a modelli predittivi.
  - In BI, questa fase coincide con la modeling task.
  - In Process Mining, si costruiscono modelli di processo. Gli artefatti progettati possono includere sistemi software, architetture informative, modelli di processo, strumenti decisionali, ecc.
- **Measure (Misurare):** Questa fase comporta la misurazione e la valutazione delle prestazioni del sistema informativo implementato attraverso l'analisi dei dati generati (come KPI o log degli eventi) per verificare il raggiungimento degli obiettivi e la soddisfazione dei requisiti. È fondamentale chiedersi se le differenze osservate sono significative per inferire nuova conoscenza, considerando la validità del metodo, l'accuratezza della raccolta dati e la rappresentatività del campione.
  - In BI, questa fase corrisponde alle analysis e evaluation tasks.
  - In Process Mining, si usa il conformance checking per misurare quanto il comportamento reale si discosti dal modello.
- **Evaluate (Valutare):** In questa fase si analizzano i risultati delle misurazioni per trarre conclusioni sull'efficacia del sistema e per identificare aree di miglioramento.
  - Questa valutazione può portare a modifiche dell'artefatto o dei requisiti, specialmente nel contesto della Design Science.

Il ciclo ODME può quindi ripetersi, portando a un continuo miglioramento del sistema.

**Design Science (DS) vs Scienza Esplicativa (ES):** Esiste una distinzione fondamentale tra queste due visioni della ricerca:

**Scienza Esplicativa (ES):** L'obiettivo principale è comprendere e spiegare i fenomeni. Se le osservazioni contraddicono l'ipotesi, l'ipotesi deve essere rifiutata o modificata. → L'enfasi è sulla verità e sulla coerenza con la realtà osservata.

**Design Science (DS):** L'obiettivo è creare artefatti (sistemi informativi, metodi, algoritmi, ecc.) che risolvano problemi e migliorino determinate situazioni. → Se le osservazioni mostrano che un artefatto non soddisfa i requisiti, si possono modificare l'artefatto stesso o i requisiti. → L'enfasi è sull'utilità e sull'efficacia dell'artefatto nel raggiungere gli obiettivi. → La DS è quindi una scienza costruttiva.

### **Rilevanza dei KPI e dell'allineamento strategico IT-business:**

I KPI sono fondamentali in ogni fase: vengono osservati per comprendere la situazione, usati come obiettivi nella progettazione, misurati nelle performance e analizzati nella valutazione. → Il metodo ODME serve anche a garantire che le decisioni IT siano coerenti con la strategia aziendale, allineando la progettazione dei sistemi informativi con gli obiettivi di business.

## Schema riassuntivo:

ODME nelle IS	Design Science vs Explanatory Science
Observe: analisi contesto, KPI, log	DS crea soluzioni utili, si adatta agli obiettivi
Design: progettazione artefatti (sistemi, modelli, interfacce)	ES→ spiega fenomeni, si adatta alla realtà osservata
Measure: valutazione tramite KPI, log, metodi quantitativi	Design Science life cycle
Evaluate: analisi efficacia, revisione sistema/requisiti	

Esporta in Fogli

### Allineamento tra artefatti e requisiti:

- Nel contesto della Design Science, se le osservazioni mostrano che un artefatto non soddisfa i requisiti, si può modificare l'artefatto o i requisiti. Questo sottolinea una relazione dinamica in cui l'artefatto è progettato per soddisfare i requisiti, ma entrambi possono evolvere in base alle osservazioni.
- In Business Intelligence, la progettazione di modelli (gli artefatti) è spesso definita in accordo con un KPI (Key Performance Indicator) per specifici bisogni aziendali. I KPI, a loro volta, sono legati agli obiettivi strategici dell'organizzazione. Pertanto, gli artefatti (modelli) dovrebbero essere in linea con i requisiti definiti dagli obiettivi e dai KPI.
- La rilevanza di un modello BI dipende dalla sua capacità di spiegare osservazioni passate e prevedere osservazioni future, il che implica un allineamento con i requisiti di analisi e gli obiettivi aziendali.
- Nel Process Mining, i modelli di processo (artefatti) possono essere creati (process discovery) o confrontati con i log degli eventi (osservazioni) per verificare la conformità, che in un certo senso valuta se l'esecuzione reale (e quindi gli artefatti che la supportano) è in linea con le regole o i requisiti definiti nel modello. Le specifiche sono "conoscenza prescrittiva" che permette alle organizzazioni di raggiungere gli obiettivi e migliorare le prestazioni. Gli artefatti, come i sistemi informativi e i modelli di processo, dovrebbero essere sviluppati in linea con queste specifiche e quindi con i requisiti.

### Allineamento tra artefatti e osservazioni:

- Il Process Mining utilizza i dati generati dai sistemi informativi (log degli eventi, osservazioni) per validare i progetti (artefatti) o per modificarli.
- Le osservazioni (KPI, log degli eventi) della realtà aziendale devono essere allineate con le specifiche e gli obiettivi. Gli artefatti (sistemi informativi, modelli) sono progettati per supportare questa realtà osservata e raggiungere gli obiettivi.
- Nel BI, i modelli di dati (artefatti) sono sviluppati in base alla disponibilità e alle proprietà dei dati (osservazioni). La scelta della prospettiva aziendale dipende anche dalla disponibilità dei dati.
- Il conformance checking confronta il modello di processo (artefatto) con i log degli eventi (osservazioni), valutando quanto il modello rifletta il comportamento reale. Discrepanze indicano la necessità di aggiornare il modello. Le tecniche di replay in Process Mining cercano di seguire il modello utilizzando le tracce dei log degli eventi (osservazioni) per

identificare deviazioni. Gli allineamenti in Process Mining forniscono un modo più dettagliato per mappare il comportamento osservato sul comportamento modellato, evidenziando le discrepanze tra artefatto (modello) e osservazioni (log).

- La qualità di un modello scoperto si misura in termini di fitness (quanto bene il modello riflette i log) e precision (quanto bene il modello descrive solo quel comportamento).

#### **Allineamento tra requisiti e osservazioni:**

- Le osservazioni della realtà devono essere allineate con le specifiche e gli obiettivi (requisiti). Eventuali dipendenze e conflitti tra gli obiettivi devono essere gestiti in modo conciso. Le specifiche (che includono gli obiettivi) devono essere confrontate con le osservazioni per verificare i risultati raggiunti e acquisire nuova conoscenza. Le performance sono misurate rispetto a indicatori chiave di prestazione (KPI). La definizione dei KPI e il monitoraggio delle prestazioni si basano sia sugli obiettivi (requisiti) che sui dati (osservazioni).
- In BI, una corretta definizione degli obiettivi richiede sia comprensione del dominio (requisiti) che dei dati (osservazioni). Le analisi qualitative dei processi possono identificare parti non necessarie o fasi a valore aggiunto, contribuendo a un allineamento più efficace tra i requisiti di efficienza e le osservazioni sull'esecuzione del processo.

In sintesi, l'allineamento tra artefatti, requisiti e osservazioni è un processo iterativo e fondamentale nella disciplina IS. La Design Science permette di adattare sia gli artefatti che i requisiti in base alle osservazioni.

# Requisiti e Artefatti nei Sistemi Informativi

Nella disciplina dei Sistemi Informativi (IS), e in particolare nei contesti di Business Intelligence (BI) e Process Mining (PM), i concetti di "requisiti" e "artefatti" sono fondamentali.

- **Requisiti:** Sono un insieme di obiettivi o specifiche che consideriamo rilevanti. Questa definizione trova riscontro nelle fonti, dove i requisiti sono spesso associati agli obiettivi aziendali e ai Key Performance Indicators (KPI), che permettono di misurare le prestazioni del business rispetto a tali obiettivi. Le specifiche sono definite come "conoscenza prescrittiva" che consente alle organizzazioni di raggiungere gli obiettivi e migliorare le prestazioni. La definizione dei KPI si basa sull'identificazione di un processo aziendale predefinito, sulla definizione dei requisiti per tale processo e sulla misurazione dei risultati rispetto agli obiettivi stabiliti.
- **Artefatto:** È qualsiasi creazione umana costruita per raggiungere un obiettivo. Nel contesto del BI e del PM, gli artefatti possono assumere diverse forme, tra cui:
  - **Modelli di Business Intelligence:** utilizzati per la rappresentazione, la comprensione e l'analisi dei processi aziendali e dei dati. Esistono diversi tipi di modelli, come i modelli di fenomeni, i modelli di dati e i modelli di teorie. La loro qualità è valutata in base a criteri come correttezza, rilevanza ed efficienza economica.
  - **Modelli di Processo:** rappresentazioni grafiche o formali dei flussi di attività aziendali, come quelli creati con BPMN o le Reti di Petri. Questi modelli possono essere creati a *design time* in accordo con i KPI per specifici bisogni aziendali.
  - **Database:** sistemi strutturati per l'archiviazione e la gestione dei dati, essenziali per tutte le attività di BI.
  - **Software:** applicazioni informatiche progettate per supportare specifiche attività aziendali e raggiungere determinati obiettivi.
  - **Event Logs:** registrazioni dettagliate degli eventi che si verificano durante l'esecuzione dei processi aziendali, fondamentali per il Process Mining.

Un artefatto può agire come requisito per un altro artefatto. Questo è un aspetto importante e si manifesta in diversi modi:

- **Linee guida per la scrittura dei requisiti:** le stesse linee guida possono essere considerate un artefatto che stabilisce i requisiti per la formulazione di requisiti specifici.
- **Vocabolario:** un vocabolario aziendale standardizzato (un artefatto) può definire i termini e i concetti che devono essere utilizzati nella specificazione dei requisiti per altri artefatti, come modelli di dati o software. Le ontologie sono un approccio popolare per la specifica di un modello per la semantica del dominio e definiscono un vocabolario di classi e relazioni.
- **Regole Aziendali (Business Rules):** queste regole (artefatti) vincolano le operazioni che si applicano a un'organizzazione e sono progettate per aiutare a raggiungere gli obiettivi, fornendo efficienza, coerenza e prevedibilità. Possono definire i requisiti di comportamento per i sistemi informativi e i processi aziendali.
- **Modelli di Processo Aziendale:** un modello di processo (un artefatto) può definire la sequenza e le condizioni delle attività necessarie per raggiungere un obiettivo. Questo modello agisce come un requisito per l'implementazione del processo stesso, per il software che lo supporta e per l'analisi delle sue prestazioni attraverso il Process Mining. Nel Process Mining, i modelli di processo possono essere confrontati con i log degli eventi per il *conformance checking*, verificando se l'esecuzione reale (e quindi gli artefatti che la supportano) è in linea con i requisiti definiti nel modello.
- **Database:** lo schema di un database (un artefatto) definisce la struttura e le relazioni dei dati, agendo come un requisito per le applicazioni software che interagiscono con esso e per i modelli di BI che lo analizzano.

- **Software:** le specifiche di un software (un artefatto) definiscono i requisiti funzionali e non funzionali per la sua realizzazione. Una volta sviluppato, il software può a sua volta imporre requisiti sui processi aziendali che supporta e sui dati che genera.
- **Event Logs:** sebbene principalmente utilizzati per l'analisi, la struttura e il contenuto di un event log (un artefatto di registrazione dei dati) possono essere definiti come requisiti per i sistemi informativi, al fine di garantire che vengano catturate le informazioni necessarie per il Process Mining e la valutazione delle prestazioni. Le linee guida per il logging sottolineano l'importanza di registrare i dati degli eventi in modo significativo e strutturato.

In conclusione, i requisiti definiscono obiettivi e specifiche; gli artefatti sono strumenti per raggiungerli. La relazione tra questi elementi è dinamica e ciclica: progettazione, implementazione, misurazione e revisione. BI e PM forniscono metodologie per garantire che gli artefatti siano coerenti con i requisiti e rispecchino le osservazioni aziendali, sostenendo il raggiungimento degli obiettivi strategici.

### Utilizzo dei dati generati da IS per validare il design attraverso il Process Mining:

Il Process Mining è un campo di ricerca che supporta l'analisi dei processi operativi attraverso i log degli eventi generati dai Sistemi Informativi (IS) durante l'esecuzione dei processi aziendali. Colma il divario tra:

- l'analisi dei processi basata su modelli tradizionali
- le tecniche di analisi data-centric come Machine Learning e Data Mining.

### Come i dati generati dagli IS convalidano il design

I dati generati dagli IS, principalmente sotto forma di log degli eventi, sono fondamentali per il Process Mining per convalidare il design in diversi modi:

#### 1. Misurazione delle Performance

- Analizza le istanze dei processi e misura le performance rispetto a KPI e obiettivi.
- I dati come timestamp e frequenza delle attività permettono di:
  - individuare colli di bottiglia
  - diagnosticare problemi di performance
  - confrontare tempo atteso vs. tempo reale di esecuzione

#### 2. Redesign dei processi

- Tramite *process discovery*, il Process Mining può ricostruire automaticamente il processo come viene effettivamente svolto.
- Il *conformance checking* confronta il modello reale con quello progettato:
  - individua deviazioni e comportamenti anomali o imprevisti.
- può guidare modifiche come:
  - riordinare le attività
  - centralizzare o ridistribuire le risorse
  - semplificare le sequenze operative

#### 3. Supporto alle decisioni (Decision Support)

- Offre *insight* oggettivi basati sui dati reali dei processi ("as-is").
- Aiuta il *decision-making* a partire da dati storici:
  - per identificare problemi
  - per evidenziare opportunità di miglioramento
  - per suggerire azioni correttive o trasformazioni

#### 4. Automazione dei processi

- Grazie alla comprensione dei workflow reali, consente:
  - l'identificazione di attività ripetitive e standardizzabili
  - la selezione di attività candidabili all'automazione

- Permette di ottenere soluzioni:
  - più veloci
  - a costo ridotto
  - con migliore impatto operativo

## **Il Process Mining e la Design Science**

Il Process Mining può essere visto come una metodologia che opera in linea con i principi della Design Science. La Design Science è focalizzata sullo sviluppo di artefatti con l'intenzione esplicita di migliorarne le performance funzionali. Essa è considerata una scienza costruttiva, in cui si mappa uno spazio di obiettivi a uno spazio di attributi.

### **Come si integra il Process Mining**

Il Process Mining utilizza i dati generati dagli IS (osservazioni) per validare il design (artefatto) dei processi aziendali e dei sistemi informativi che li supportano. Se le osservazioni (dati dei log) mostrano che il processo non è in linea con gli obiettivi o le specifiche (requisiti), il Process Mining fornisce le informazioni necessarie per modificare l'artefatto (il processo o il sistema informativo) o per ricalibrare i requisiti. Questo ciclo iterativo di osservazione, design, misurazione e valutazione (ODME) è centrale nella Design Science.

### **Il ciclo iterativo ODME:**

- Osservazione: raccolta dei dati dai log generati dai IS
- Design: definizione/modifica dell'artefatto (es. modello di processo)
- Misurazione: valutazione tramite KPI e log
- Valutazione: analisi delle performance, verifica della conformità

Se i log mostrano discrepanze tra esecuzione reale e modello:

- si può modificare il design
- oppure ricalibrare i requisiti

### **Difficoltà epistemologiche:**

Come ogni attività di modellazione, anche il Process Mining eredita delle difficoltà epistemologiche.

- Un modello è una rappresentazione della realtà, e non è detto che rappresenti completamente tutto il comportamento possibile del sistema.
- I log degli eventi contengono esempi di comportamento, ma non possiamo sapere se coprono interamente il comportamento che il modello dovrebbe rappresentare.
- Inoltre, la scelta del modello di processo e delle tecniche di analisi può influenzare l'interpretazione dei risultati.

### **Connessioni con altri metodi, strumenti e campi**

Il Process Mining lavora in connessione con molti altri metodi, strumenti e campi:

- **Business Intelligence (BI):** Il Process Mining è complementare al BI. Mentre il BI si concentra su dashboard e reporting, il Process Mining fornisce insight sui processi aziendali. Entrambi utilizzano i dati per supportare le decisioni aziendali.
- **Data Mining e Machine Learning:** Il Process Mining si basa su tecniche di Data Mining per:
  - estrarre pattern dai log degli eventi
  - Può usare algoritmi di Machine Learning per:
    - predizioni

- e supporto operativo (es. suggerimenti automatici)
- **Business Process Management (BPM):**
  - Il Process Mining fornisce una visione oggettiva dei processi "as-is", che può essere utilizzata per migliorare i processi definiti nel BPM.
  - Il conformance checking verifica se l'esecuzione reale è conforme ai modelli di processo definiti nel BPM.
- **Data Science:**
  - Il Process Mining è considerato una sotto-disciplina della Data Science che si concentra specificamente sull'analisi dei dati di processo.
  - I data scientist possono utilizzare le tecniche di Process Mining per ottenere valore dai dati degli eventi e rispondere a domande sul:
    - "cosa è successo",
    - "perché è successo",
    - "cosa succederà" e
    - "cosa è meglio che succeda" in relazione ai processi.
- **Strumenti di Process Mining:** Esistono numerosi strumenti commerciali e open-source (come PM4PY e ProM) che supportano le diverse tecniche di Process Mining, inclusi la process discovery, il conformance checking e l'analisi delle performance.

In conclusione, il Process Mining utilizza i dati generati dai Sistemi Informativi per fornire una validazione basata sui fatti del design dei processi, consentendo di misurare le performance, supportare il redesign, l'automazione e il decision making. La sua natura iterativa e la capacità di utilizzare le osservazioni per informare il design lo rendono coerente con i principi della Design Science, pur ereditando le sfide legate alla modellazione della realtà. La sua forte connessione con altri campi come il BI, il Data Mining, il BPM e la Data Science ne sottolinea la sua natura interdisciplinare e la sua importanza nell'analisi e nel miglioramento dei processi aziendali.

## **SDLC – Systems Development Life Cycle (o IS Development Life Cycle)**

Lo SDLC è un modello concettuale usato per guidare e strutturare tutte le fasi coinvolte nello sviluppo di un Sistema Informativo (IS). È impiegato in Project Management per gestire la complessità, adottando metodologie e strumenti adeguati.

### **Fasi tipiche del ciclo di vita SDLC**

Anche se i modelli possono variare leggermente, le fasi classiche sono:

È fondamentale per la progettazione e gestione efficace di Sistemi Informativi. Prevede fasi ben definite: Analisi, Design, Implementazione, Testing, Valutazione, Manutenzione. È flessibile e iterativo, con collegamenti diretti a discipline come BI, BPM, Process Mining, e Design Science per supportare decisioni e miglioramenti continui.

1. **Analisi dei requisiti (Requirements Analysis)**
  - Comprendere i bisogni del business.
  - Raccogliere e documentare i requisiti funzionali e non funzionali.
  - Coinvolge stakeholder, utenti e analisti.
2. **Progettazione (Design)**
  - Si definisce l'architettura logica e tecnica del sistema.
  - Include design dell'interfaccia utente, del database e dei flussi di processo.
  - In contesto Design Science: si può adattare il design se i requisiti non sono soddisfatti.
3. **Implementazione (Implementation)**
  - Il sistema viene codificato e costruito sulla base del design.



- Si usano linguaggi di programmazione, tecnologie e strumenti appropriati.
  - In ambito BI o BPM, può riferirsi all'implementazione di soluzioni analitiche o processuali.
4. **Testing**
    - Verifica che il sistema funzioni correttamente e sia conforme ai requisiti.
    - Include test funzionali, di integrazione, di sistema e di accettazione.
    - In ambito Process Mining, il conformance checking è un'analogia: verifica che i processi reali aderiscano al modello.
  5. **Valutazione (Evaluation)**
    - Analisi dei risultati rispetto agli obiettivi aziendali.
    - Si controlla l'efficacia, l'efficienza, la qualità.
    - Nell'iMine format, questa fase è chiamata "Evaluation and reporting task".
  6. **Manutenzione e aggiornamento (Maintenance)**
    - Dopo il rilascio, il sistema può necessitare aggiornamenti, patch, miglioramenti.
    - In BPM, questo equivale alla fase di monitoraggio e diagnosi.
  7. **Iteratività**
    - Lo SDLC è iterativo: nuove analisi e insight (es. da Process Mining o BI) possono innescare nuovi cicli di sviluppo o redesign.

## BPM LIFE CYCLE

Il ciclo di vita del Business Process Management (BPM) è un approccio iterativo che supporta le organizzazioni nella pianificazione, gestione e miglioramento continuo delle loro attività. Si compone di diverse fasi:

- **Design:** in questa fase, il processo viene progettato. Si definiscono le attività, le sequenze operative, i ruoli coinvolti e gli obiettivi del processo.
- **Implementazione/Configurazione:** il modello progettato viene trasformato in un sistema funzionante. Ciò include l'impostazione tecnica e organizzativa necessaria per rendere operativo il processo.
- **Esecuzione e Monitoraggio:** il processo viene messo in esecuzione e monitorato nel tempo. Si osservano le prestazioni del processo durante la sua attività concreta.
- **Diagnosi e Requisiti:** in questa fase si valutano le prestazioni del processo e si identificano eventuali problemi o nuove esigenze dovute a cambiamenti nel contesto aziendale. Quando emergono criticità o opportunità di miglioramento, si avvia una nuova iterazione del ciclo, ripartendo dalla fase di redesign.
- **Redesign:** se la diagnosi evidenzia carenze o nuove necessità, il processo viene riprogettato per migliorarne efficienza, efficacia o adeguatezza al contesto.

L'utilizzo dei dati reali derivanti dall'esecuzione del processo permette di basare l'analisi su informazioni concrete. Questo consente una valutazione oggettiva delle prestazioni e delle deviazioni rispetto al modello atteso, migliorando la qualità e la precisione delle decisioni di cambiamento.

## Analisi delle attività specifiche

- **Identificazione del processo:** è la fase iniziale in cui si individuano le regole che definiscono le relazioni tra gli eventi del processo. Serve a delineare cosa sarà oggetto di analisi e come è strutturato il processo.
- **Discovery:** consiste nel ricavare automaticamente un modello di processo partendo dai dati di esecuzione (eventi registrati). Si ottiene una rappresentazione reale del processo così come si è svolto, senza fare affidamento su ipotesi iniziali.
- **Analisi:** si studiano le caratteristiche del processo per rispondere a domande sul suo funzionamento. L'analisi può essere qualitativa (es. quali attività non hanno un responsabile

assegnato) o quantitativa (es. qual è la durata media di una determinata attività). L'analisi può basarsi solo sul modello oppure includere i dati reali del processo.

- **Redesign:** si modificano i processi sulla base delle analisi effettuate, con l'obiettivo di migliorare l'efficienza, ridurre i costi, soddisfare nuovi requisiti o adattarsi a cambiamenti ambientali.
- **Implementazione:** si realizzano concretamente le modifiche definite nella fase di redesign. Questo può includere l'automazione, la modifica dei flussi di lavoro, o l'aggiornamento dei sistemi informativi.
- **Monitoraggio:** consiste nell'osservazione continua del processo in esecuzione per verificare che funzioni come previsto. Vengono tracciati indicatori di prestazione e identificati eventuali problemi o ritardi.
- **Concealing discovery:** questo termine non è comunemente utilizzato in questo contesto. Potrebbe riferirsi a strategie per nascondere o limitare la visibilità di informazioni emerse durante l'analisi, ma non ha un significato standard o una definizione formalizzata nel ciclo di vita dei processi.

# Design dei Sistemi Informativi: Specifiche e Strumenti

Il design dei Sistemi Informativi (SI) è un processo critico che richiede specifiche dettagliate per documentare l'intero ciclo di vita dello sviluppo. Queste specifiche non solo permettono di identificare, descrivere, prescrivere e verificare ciò che si intende implementare, ma forniscono anche input cruciali per i modelli predittivi, aiutando a definire le tendenze osservate.

Le specifiche più importanti includono:

- **Obiettivi (Goals)**
- **Regole di Business (Business Rules)**
- **Modelli dei Processi di Business (Business Process Models)**
- **Modelli del Flusso di Dati (Data Flow Models)**

Altre specifiche rilevanti sono report, questionari, log di eventi, vocabolari, direttive, ontologie, requisiti e regolamenti. Per essere efficaci, tutte le specifiche devono essere costantemente confrontate con le osservazioni reali, al fine di verificarne il raggiungimento e acquisire nuova conoscenza.

## 1. Obiettivi (Goals)

Gli obiettivi sono lo scopo finale dell'organizzazione e un elemento cruciale nella progettazione dei SI. L'acquisizione di conoscenza è utile solo se supporta il raggiungimento o la verifica di tali obiettivi. I vantaggi di definire chiaramente gli obiettivi sono molteplici:

- Focalizzazione sulle attività più rilevanti
- Comprensione delle interconnessioni tra le diverse parti dell'organizzazione
- Esplorazione di alternative strategiche
- Valutazione accurata delle performance

Nonostante la loro importanza, molte metodologie attuali offrono scarso supporto esplicito per l'elicitazione e la rappresentazione degli obiettivi.

La definizione di un obiettivo implica la misurazione di una o più "dimensioni misurabili". Se ci sono più dimensioni, è necessaria una funzione di aggregazione per sintetizzarle in un unico valore, che rappresenta l'indicatore. I valori che un indicatore può assumere sono spesso chiamati "livelli". L'indicatore può essere direttamente collegato a un obiettivo che specifica il livello da raggiungere o il miglioramento atteso.

Nel contesto dei Sistemi Informativi, in particolare nella Business Intelligence (BI), gli obiettivi di analisi sono il punto di partenza per lo sviluppo di applicazioni e spesso vengono formulati tramite i Key Performance Indicators (KPI). Gli obiettivi analitici definiscono il tipo di analisi da condurre (descrittiva, predittiva o di comprensione) e permettono una specifica formale del target in relazione a fattori influenti. Nei progetti di Process Mining orientati agli obiettivi, questi vengono formulati nella prima fase del ciclo di vita.

## 2. Regole di Business (Business Rules)

Le regole di business (BR) sono vincoli che si applicano alle operazioni di un'organizzazione. Sono progettate per aiutare l'organizzazione a raggiungere i suoi obiettivi e a garantire efficienza, coerenza e prevedibilità. Secondo il Business Rules Group, le BR possono:

- **Definire un vocabolario di termini** adottati nell'organizzazione. Esempi includono definizioni di "cliente" (uno che acquista un bene o servizio), "ordine del cliente" (il cliente

trasmette l'ordine all'azienda), "ordine fermo" (ordine finale), "lettera di intenti" (ordine non finale).

- **Descrivere fatti che possono essere osservati** nell'organizzazione. Ad esempio: "L'ordine 123 è finale", "La lettera di intenti 12 è stata ricevuta Giovedì 15 Ottobre", "Lo studente con matricola 1234 si è iscritto al corso di Process Mining".
- **Vincolare il comportamento** dell'organizzazione. Alcuni esempi sono: "Ogni ordine deve essere archiviato nell'archivio", "Ogni studente che si iscrive a un corso deve pagare 300\$ per il corso", "Ogni scuola deve fornire tutto il materiale didattico per una classe a ogni studente della classe il primo giorno di lezione", "Ogni persona che entra negli edifici dell'Università deve esibire il Green Pass".
- **Spiegare come la conoscenza può essere derivata o trasformata.** Esempi: "Uno studente è una persona", "Un professore è una persona", "Un dipendente è una persona", "OTH è un'Università", "Qualsiasi edificio utilizzato per le attività dell'Università è un edificio dell'Università".

Le BR possono essere documentate come frasi in linguaggio naturale, linguaggi controllati o strutture in un modello grafico.

Il **Semantics of Business Vocabulary and Business Rules (SBVR™)** è uno standard OMG per specificare le regole di business. È un linguaggio controllato, ovvero un sottoinsieme dell'inglese standard con sintassi e semantica ristrette, descritto da un piccolo insieme di regole di costruzione e interpretazione. L'SBVR consente di esprimere termini, verbi, individui, definizioni, quantificatori e condizioni. Ad esempio:

- **Quantificatori:** "ogni noleggio ha almeno un guidatore", "è obbligatorio che ogni noleggio non abbia più di 4 guidatori", "è consentito che un ordine sia pagato in contanti".
- **Condizioni:** "se una restituzione del noleggio è in ritardo di più di 4 ore, si applica una multa".

Le regole di business possono essere **aletiche** o **deontiche**:

- **Aletiche:** Indicano come è il mondo. Se violate, segnalano una disfunzione o un'anomalia nell'operazione dell'organizzazione.
- **Deontiche:** Indicano come il mondo dovrebbe essere. Se violate, segnalano un obiettivo mancato o una performance non raggiunta.

Un linguaggio controllato non progettato correttamente può facilmente diventare intrattabile dal punto di vista della risoluzione formale delle sue dichiarazioni. Sottoinsiemi del linguaggio possono essere mappati alla logica del primo ordine o ad altri linguaggi formali della famiglia della Logica Temporale e della Logica Modale.

### 3. Modelli dei Processi di Business (Business Process Models)

Qualsiasi organizzazione organizza il proprio lavoro identificando i propri obiettivi e le attività necessarie per raggiungerli. I metodi per identificare e rappresentare tali attività possono variare nel grado di formalità. In alcuni casi, ci si può riferire all'uso e alla tradizione, ma solitamente esiste una documentazione. Le organizzazioni più strutturate hanno sviluppato la nozione di Business Process. Le organizzazioni moderne adottano sempre più il Business Process Modelling (BPM), ovvero tecniche che supportano la descrizione, la prescrizione e la spiegazione del processo delle attività.

La **Business Process Model and Notation (BPMN)** è uno standard sviluppato dall'OMG. La versione 2.0 di BPMN è stata rilasciata a gennaio 2011, e una specifica ISO (ISO 19510) pubblicata nel 2014 riflette questo standard. BPMN è uno strumento per:

- Definire il workflow
- Identificare i messaggi da scambiare
- Identificare gli eventi da monitorare
- Identificare le responsabilità
- Identificare le attività che possono essere automatizzate

Per fornire semantica formale al Process Mining, si utilizzano modelli Turing-completi come le Reti di Petri o la logica temporale. Le Reti di Petri sono un linguaggio di modellazione matematica per descrivere sistemi distribuiti e concorrenti.

Nel ciclo di vita del BPM, i modelli di processo sono centrali nelle fasi di (ri)design e configurazione/implementazione. Tuttavia, l'efficacia è limitata se i modelli non sono allineati con la realtà. Il Process Mining mira a scoprire, monitorare e migliorare i processi reali estraendo conoscenza dagli event log, fornendo un collegamento tra processi effettivi e modelli.

#### 4. Modelli del Flusso di Dati (Data Flow Models) e Data Flow Diagrams (DFD)

I **Flussi di Dati (Data Flows)** si trovano all'intersezione tra la gestione dei processi (Process Management) e i modelli di infrastruttura, che sono modelli utilizzati per progettare componenti e sistemi software. Spesso, i modelli di processo di business descrivono i processi senza suggerire come vengano condotti. I modelli di infrastruttura, invece, includono informazioni su come i processi sono implementati. In pratica, i Flussi di Dati possono essere utilizzati per specifiche astratte, mentre i modelli BPM possono essere arricchiti con dettagli sull'elaborazione delle informazioni, sebbene quest'ultimo sia meno comune.

I dati e il loro flusso sono centrali per la Business Intelligence e il Process Mining. Il data provisioning include la raccolta, l'estrazione, la trasformazione e l'integrazione dei dati per l'analisi. Diverse viste sui processi (eventi, stato, trasversale) generano dati che fungono da input per la BI. Nel Process Mining, gli event log registrano le attività e il loro ordine, tracciando il flusso degli eventi e sono fondamentali per le tecniche di scoperta dei processi, verifica della conformità e miglioramento.

I **Data Flow Diagrams (DFD)** sono strumenti grafici per visualizzare il flusso di informazioni all'interno di un sistema. Mostrano come i dati vengono elaborati e trasformati mentre si spostano da una fonte esterna a una destinazione, attraverso processi e depositi di dati.

Gli **elementi grafici** di un DFD includono:

- **Processo:** Rappresenta un'attività che trasforma i dati (graficamente un cerchio o un rettangolo con angoli arrotondati e un numero ID).
- **Deposito di Dati (Data Store):** Rappresenta un luogo di archiviazione dei dati (graficamente due linee parallele o un rettangolo aperto da un lato).
- **Entità Esterna (External Entity):** Rappresenta fonti o destinazioni esterne di dati (graficamente un rettangolo).
- **Flusso di Dati:** Rappresenta il movimento dei dati tra gli elementi (graficamente una freccia).

La **decomposizione** è il processo di modellazione del sistema e dei suoi componenti con livelli di dettaglio crescenti. Il **bilanciamento** assicura che le informazioni presentate a un livello del DFD siano rappresentate accuratamente nel DFD del livello successivo. I livelli tipici di un DFD sono:

- **Diagramma di Contesto:** Mostra il contesto in cui si inserisce il processo di business, rappresentando l'intero processo come un unico elemento e tutte le entità esterne che ricevono o contribuiscono informazioni al sistema.
- **Diagramma di Livello 0:** Mostra tutti i processi che compongono il sistema complessivo, come le informazioni si muovono da e verso ciascun processo, e aggiunge i depositi di dati.
- **Diagramma di Livello 1:** Mostra tutti i processi che compongono un singolo processo del diagramma di livello 0, come le informazioni si muovono tra questi processi e dettaglia ulteriormente il contenuto del processo di livello superiore. I diagrammi di Livello 1 potrebbero non essere necessari per tutti i processi di livello 0.
- **Diagramma di Livello 2:** Mostra tutti i processi che compongono un singolo processo del diagramma di livello 1, e come le informazioni si muovono tra questi processi. Anche i diagrammi di Livello 2 potrebbero non essere necessari per tutti i processi di livello 1.

Una corretta numerazione di ogni processo aiuta il progettista a capire dove il processo si inserisce nel sistema complessivo.

Aspetti **tecnologici** importanti nella progettazione di un Flusso di Dati riguardano la tecnologia adottata per implementarlo, in particolare gli aspetti che impattano sulle attività di elaborazione dei dati. Questi includono:

- **Stream vs Batch processing:** elaborazione continua vs elaborazione a lotti.
- **ETL vs ELT:** Extract, Transform, Load vs Extract, Load, Transform.
- **Consistency vs Availability:** compromessi tra coerenza dei dati e disponibilità del sistema.
- **Pseudonymization vs Anonymization:** tecniche di protezione dei dati.

## Specifiche e Knowledge Uplift Trail (KUT)

Le specifiche esprimono principalmente Conoscenza Descrittiva o Prescrittiva, definendo come un'organizzazione deve funzionare. Possono anche esprimere Conoscenza Predittiva supportando:

- **Analisi di Conformità (Conformance Analysis):** Il problema è descritto dall'implicazione: Fatti, Conoscenza Estesa → Vincoli.
- **Previsione delle Performance (Performance Prediction):** Il problema è descritto dalla funzione:  $f(\text{Fatti, Conoscenza Estesa}) = \text{nuovi Fatti}$ .

Qui, i **Fatti** includono regole di business, modelli di processo, flussi di dati, report, questionari, log di eventi. La **Conoscenza Estesa** comprende regole di business, vocabolari, direttive, ontologie. I **Vincoli** sono modelli di valore, regole di business, modelli di processo, flussi di dati, requisiti, regolamenti, direttive.

## Case Study: Dati Farmaceutici

Un caso studio riguarda EA, un'azienda che pubblica informazioni su farmaci e dispositivi medici, raccolte ogni quattro mesi in un manuale venduto a medici e professionisti sanitari. EA sta modificando il suo modello, introducendo la possibilità per i clienti di pagare una quota annuale e ricevere aggiornamenti ogni volta che nuove informazioni su un singolo articolo sono disponibili. Per realizzare ciò, EA sta distribuendo le sue informazioni sul web, limitando l'accesso agli utenti registrati. Per aumentare la diffusione dei suoi prodotti, EA sta anche riducendo le tariffe per accessi ristretti a specifiche porzioni della knowledge base. Inoltre, EA sta raccogliendo informazioni statistiche sulle query degli utenti al fine di costruire una knowledge base delle preferenze nelle pratiche mediche. Questa nuova knowledge base può essere utilizzata per vendere informazioni statistiche ai produttori di servizi sanitari. Questa conoscenza è utile anche per definire direttive per il ramo commerciale dell'azienda, al fine di raccomandare buone pratiche per l'approccio al cliente.

Le domande chiave per questo caso studio sono:

- Quali obiettivi vuole raggiungere EA?
- Quale conoscenza viene utilizzata per raggiungere l'obiettivo?
- Quali passaggi consentono di elevare questa conoscenza (uplift)?
- Quali informazioni consentono in ogni passaggio di acquisire questa conoscenza?
- È possibile codificare parte di queste informazioni utilizzando Regole di Business, Modelli di Processo o Modelli di Flusso di Dati?

Per rispondere a queste domande, è possibile:

- Evidenziare gli obiettivi utilizzando un modello di valore.
- Sviluppare un Knowledge Uplift Trail (KUT).
- Verificare che il KUT consenta all'organizzazione di raggiungere i suoi obiettivi.
- Se pertinente, fornire esempi utilizzando un linguaggio di specifica introdotto in classe.

Gli obiettivi del caso studio includono la vendita di informazioni all'industria farmaceutica e la raccolta di informazioni sul mercato dei professionisti.

Un esempio di KUT per questo caso studio potrebbe essere:

Input	Conoscenza Acquisita	Output	Tipo Analisi/
<b>Step 1:</b> Abbonamenti all'Online Handbook per categoria e paese	Frequenza degli abbonamenti	Abbonamenti per paese e categoria	Descrittivo
<b>Step 2:</b> Step 1	Verifica delle entrate del modello di abbonamento per	Confronto dei ricavi con modelli diversi	Descrittivo
<b>Step 3:</b> Step 1	Piani di marketing per paese	Eventi di marketing	Prescrittivo
<b>Step 4:</b> Step 1, Step 2	Previsione dei ricavi per l'anno	Previsione dei	Predittivo

## Modellare il Valore con il Framework e<sup>3</sup>value

Il framework e<sup>3</sup>value è una metodologia per modellare come il valore economico viene creato, scambiato e percepito in una rete di attori. Estende gli strumenti dell'informatica per adattarsi alle dinamiche incerte del business, rappresentando formalmente il valore economico generato e trasferito. È possibile scaricare un software di supporto gratuito per e<sup>3</sup>value da [e3value.few.vu.nl](http://e3value.few.vu.nl).

### Modellare una Costellazione di Valore (Eliciting a constellation)

Il framework utilizza domande chiave per costruire una "costellazione" di attori e scambi di valore:

- **Chi sono gli attori coinvolti?** Si identificano gli attori (imprese e clienti finali) coinvolti. Gli attori sono entità economicamente indipendenti (es. imprese e consumatori, unità di business responsabili di profitti e perdite). Graficamente, sono rappresentati da un rettangolo con il nome dell'impresa o del ruolo.
- **Cosa scambiano e cosa chiedono in cambio?** Si determina cosa trasferiscono gli attori l'uno all'altro in termini di valore economico e cosa richiedono in cambio. Gli oggetti di valore (Value Objects) sono servizi, beni, denaro o anche un'esperienza, che hanno valore economico per almeno uno degli attori coinvolti nel modello di valore. Essi modellano cose di valore economico che possono essere osservate (es. Denaro, Musica, Auto, Elettricità). La valutazione stessa è soggettiva.

- **Perché avvengono questi scambi?** Si identifica quali domande governano il modello e le motivazioni dietro i trasferimenti di valore.
- **Quali attività svolgono gli attori?** Si descrivono le attività che gli attori svolgono per produrre/consumare valore.

### **Analizzare una Costellazione (Analysing properties of a constellation)**

Dopo la costruzione, si analizzano diverse proprietà:

- **Sostenibilità economica** per tutti gli attori.
- **Fattibilità tecnica.**

### **Costrutti principali di e<sup>3</sup>value**

- **Actor:** Modella un'entità economicamente indipendente.
- **Value Object:** Rappresenta un bene, servizio, denaro o esperienza di valore economico.
- **Value Port:** Modella la fornitura o la richiesta di oggetti di valore da o verso l'ambiente dell'attore. Implica un cambiamento di proprietà o di diritti. È usato per astrarre i processi di business interni. Esempi: richiesta (in-port) di denaro, offerta (out-port) di un bene.
- **Value Offering:** Raggruppa porte con la stessa direzione e modella il "bundling" (pacchettizzazione di valori), in cui il valore è solo in combinazione. Graficamente è rappresentato in modo specifico.
- **Value Interface:** Raggruppa offerte di valore in entrata e in uscita e modella la reciprocità economica. Esempio: un'interfaccia di valore può includere "musica + pagamento + accesso online + pagamento". Graficamente è rappresentato in modo specifico.
- **Value Transfer:** Connette due porte di valore tra loro e modella uno o più potenziali scambi di oggetti di valore, mostrando quali attori sono disposti a trasferire oggetti di valore l'uno all'altro. Graficamente, è una connessione tra un in-port (bene) e un out-port (pagamento).
- **Market Segment:** Scomponi un mercato (composto da attori) in segmenti che condividono proprietà comuni. Modella che un certo numero di attori assegna valore economico agli oggetti nello stesso modo.
- **Value Activity:** Attività interne all'attore per produrre o consumare il value object.
- **Composed Actor (Actor Composition):** Modella che gli attori offrono qualcosa di valore economico congiuntamente come una partnership (non ownership).

### **Dependency Path**

- Le interfacce di valore tra diversi attori/attività sono correlate tramite trasferimenti di valore.
- Le interfacce di valore dello stesso attore/attività sono correlate da percorsi di dipendenza.
- Lo scopo è rendere i modelli di valore "computabili".
- Possono includere stimoli di inizio e fine, con operatori logici OR e AND.

### **AS-IS vs TO-BE**

- Confrontare più versioni di un modello è utile per verificare l'impatto di specifiche scelte di design.
- **AS-IS:** documenta ciò che è attualmente in atto.
- **TO-BE:** documenta ciò che viene proposto.
- Lo studio dei modelli e delle specifiche aiuta le organizzazioni a ottimizzare i propri processi per migliori performance, maggiore efficienza e risultati migliorati



# Performance Measurement: Fondamenti, Metodologie e Integrazione

## Introduzione

La **Performance Measurement** rappresenta un'attività gestionale cruciale per lo sviluppo della conoscenza organizzativa. Attraverso un processo sistematico di reporting delle prestazioni, essa fornisce le informazioni essenziali per comprendere lo stato attuale dell'organizzazione, monitorare il progresso verso gli obiettivi strategici e identificare aree di potenziale miglioramento. La misurazione della performance supporta il processo decisionale a vari livelli aziendali, fornendo una base oggettiva per l'azione.

## Definizione

La misurazione della performance è definita come il processo sistematico di raccolta, analisi e reporting di informazioni quantitative e qualitative relative alle prestazioni di un'organizzazione, dei suoi processi, delle sue unità operative o dei singoli individui.

L'obiettivo primario è valutare il grado di raggiungimento degli obiettivi prefissati e fornire insight per il miglioramento continuo e il supporto alle decisioni strategiche e operative.

## Metodologie Principali per la Misurazione della Performance

Due delle metodologie più diffuse e influenti nella misurazione della performance sono la **Balanced Scorecard (BSC)** e i **Key Performance Indicators (KPI)**.

### 1. Balanced Scorecard (BSC)

Originariamente concepita come strumento di reporting direzionale, la Balanced Scorecard si è evoluta in un framework strategico completo che permette di definire, misurare e monitorare gli obiettivi strategici di un'organizzazione attraverso quattro prospettive interconnesse:

- **Finanziaria:** misura le performance finanziarie e la creazione di valore per gli azionisti.
- **Clienti:** valuta la soddisfazione, la fidelizzazione e l'acquisizione dei clienti.
- **Processi interni:** analizza l'efficienza e l'efficacia dei processi chiave che guidano la creazione di valore per i clienti e l'organizzazione.
- **Apprendimento e crescita:** considera le capacità organizzative, l'innovazione, la formazione e la cultura aziendale necessarie per sostenere il miglioramento continuo e la crescita a lungo termine.

### Evoluzione del modello BSC

Il modello BSC ha subito diverse evoluzioni, culminando in una terza generazione più sofisticata, strutturata in:

- **Dichiarazione di destinazione:** vision e mission aziendali che definiscono l'aspirazione e lo scopo dell'organizzazione.

- **Modello di collegamento strategico (Strategy Map):** una rappresentazione visiva delle relazioni causa-effetto tra gli obiettivi strategici nelle diverse prospettive, evidenziando come le azioni in un'area influenzano i risultati in altre.
- **Obiettivi strategici:** dichiarazioni qualitative di ciò che l'organizzazione intende raggiungere.
- **Misure e target quantitativi:** indicatori specifici (KPI) e valori target definiti per monitorare il progresso verso gli obiettivi strategici.

### **Integrazione con la Business Intelligence (BI)**

Nel contesto della BI, la Balanced Scorecard rappresenta spesso il prodotto finale di un processo analitico. I target definiti vengono monitorati attraverso **data warehouse** che centralizzano i dati aziendali e strumenti di visualizzazione che rendono le informazioni facilmente interpretabili. Questo offre una visione unificata e misurabile dell'esecuzione della strategia aziendale.

## **2. Key Performance Indicators (KPI)**

I **KPI** sono quantità misurabili che forniscono informazioni sul livello di efficacia e di efficienza con cui un'organizzazione sta perseguendo specifici obiettivi.

Essi rappresentano un collegamento diretto tra le attività operative e gli obiettivi strategici. I KPI sono strumenti fondamentali per:

- Monitorare la performance di specifici processi, attività o aree funzionali.
- Valutare i risultati conseguiti rispetto agli obiettivi prefissati e ai benchmark.
- Supportare il processo decisionale fornendo dati oggettivi per identificare problemi, opportunità e guidare le strategie di miglioramento.

### **Tipologie di KPI**

I KPI possono essere classificati in diverse tipologie in base alle loro caratteristiche:

- **Quantitativi:** basati su dati numerici e misurabili.
- **Pratici:** misurabili utilizzando strumenti e dati disponibili all'interno dell'organizzazione.
- **Direzionali:** indicano la tendenza della performance (miglioramento o peggioramento).
- **Azionabili:** forniscono insight su aree in cui è possibile intervenire per migliorare i risultati.
- **Finanziari:** legati ai risultati economici e alla performance finanziaria dell'organizzazione.

### **Ruolo dei KPI nella Business Intelligence**

Nell'ambito della BI, i KPI sono spesso il punto di partenza dell'analisi. I sistemi analitici mirano a comprendere come i diversi fattori influenti (driver di performance) impattino sui KPI.

L'obiettivo è identificare le leve su cui agire per migliorare i risultati e supportare strategie decisionali basate sui dati (data-driven).

# Integrazione con Business Intelligence e Process Mining

## 1. Business Intelligence (BI) e Corporate Performance Management (CPM)

La misurazione della performance è intrinsecamente legata al **Business Performance Management (BPM)**, spesso denominato anche **Corporate Performance Management (CPM)**.

Entrambi gli approcci si concentrano sulla gestione della performance aziendale, integrando la strategia con le operazioni.

- La **Business Intelligence** si focalizza sull'analisi di dati storici e attuali per generare insight e conoscenza sul business. Fornisce le informazioni necessarie per comprendere *cosa è successo e perché*.
- Il **Corporate Performance Management** utilizza questi insight forniti dalla BI per guidare e migliorare l'esecuzione strategica. Si concentra su *cosa fare* per raggiungere gli obiettivi futuri.

BI e CPM devono coesistere e supportarsi reciprocamente. La BI fornisce i dati e le analisi che alimentano i processi di CPM, mentre il CPM definisce le esigenze informative e gli obiettivi che guidano le attività di BI.

## 2. Process Mining

Anche nel **Process Mining**, la misurazione della performance riveste un ruolo centrale. I KPI rappresentano lo strumento principale per valutare l'efficacia e l'efficienza dei processi aziendali. Le metriche chiave analizzate nel Process Mining includono:

- **Tempo:** *lead time* (tempo totale del processo), *service time* (tempo di esecuzione di un'attività), *waiting time* (tempo di attesa tra le attività).
- **Costo:** costi diretti e indiretti associati all'esecuzione del processo.
- **Qualità:** tasso di conformità (rispetto delle regole), numero di difetti o rilavorazioni, customer satisfaction (misurata attraverso indicatori legati al processo).

Un progetto di process mining è spesso **goal-oriented**, ovvero mira a migliorare un processo specifico rispetto a KPI predefiniti.

L'analisi della performance, attraverso la visualizzazione dei colli di bottiglia, delle inefficienze e delle deviazioni dai percorsi ottimali, è una delle funzionalità principali degli strumenti commerciali di process mining.

## Sfide nella Misurazione della Performance

L'implementazione efficace di sistemi di misurazione della performance può incontrare diverse sfide, tra cui:

- **Definizione di KPI pertinenti:** identificare indicatori che realmente riflettano gli obiettivi strategici e le performance chiave.
- **Raccolta e analisi di dati affidabili:** garantire la qualità, l'accuratezza e la tempestività dei dati utilizzati per il calcolo dei KPI.

- **Resistenza al cambiamento:** superare la potenziale opposizione da parte dei dipendenti che potrebbero percepire la misurazione come uno strumento di controllo punitivo.
- **Allineamento delle misure con la strategia:** assicurare che i KPI e gli obiettivi di performance siano direttamente collegati alla strategia aziendale complessiva.
- **Eccessiva focalizzazione sui numeri:** evitare che la misurazione porti a un'attenzione eccessiva ai soli dati quantitativi, trascurando aspetti qualitativi importanti.

## Best Practices per una Misurazione Efficace della Performance

Per superare le sfide e garantire l'efficacia della misurazione della performance, è fondamentale adottare alcune best practices:

- **Coinvolgimento degli stakeholder:** includere le parti interessate nella definizione dei KPI per garantirne la rilevanza e l'accettazione.
- **KPI SMART:** assicurarsi che i KPI siano *Specifici, Misurabili, Raggiungibili, Rilevanti e Temporizzati*.
- **Revisione periodica dei KPI:** valutare regolarmente la pertinenza dei KPI e modificarli o aggiornarli in base all'evoluzione della strategia e del contesto aziendale.
- **Utilizzo dei dati per l'azione:** impiegare i risultati della misurazione per identificare aree di miglioramento e implementare azioni concrete.
- **Comunicazione trasparente:** condividere i risultati della performance con i dipendenti per promuovere la consapevolezza e la responsabilità.
- **Bilanciamento delle misure:** utilizzare un mix di indicatori finanziari e non finanziari per ottenere una visione completa della performance.
- **Creazione di una cultura della performance:** promuovere una mentalità in cui la misurazione è vista come uno strumento di apprendimento e miglioramento continuo.

## Ruolo della Tecnologia a Supporto della Performance Measurement

Diverse tecnologie svolgono un ruolo cruciale nel facilitare la misurazione e la gestione della performance:

- **Sistemi ERP (Enterprise Resource Planning):** forniscono una base dati integrata per la raccolta di informazioni operative e finanziarie.
- **Data Warehouse:** centralizzano e strutturano i dati provenienti da diverse fonti per l'analisi e il reporting.
- **Software di Business Intelligence:** offrono strumenti per l'analisi, la visualizzazione e la creazione di report sui KPI e sulla performance complessiva.

- **Piattaforme di CPM (Corporate Performance Management):** forniscono funzionalità specifiche per la definizione di obiettivi, la creazione di scorecard, il budgeting e la pianificazione.
- **Strumenti di Process Mining:** permettono di analizzare i log degli eventi per visualizzare i processi, identificare inefficienze e misurare la performance a livello di processo.
- **Strumenti di analisi predittiva:** utilizzano modelli statistici e algoritmi di machine learning per prevedere le performance future sulla base dei dati storici.

## Conclusioni

La **Performance Measurement**, supportata da metodologie come la Balanced Scorecard e i Key Performance Indicators, rappresenta un processo fondamentale per acquisire una profonda comprensione dell'andamento organizzativo.

Essa consente di monitorare sia la performance strategica che quella operativa, fornendo input essenziali per:

- La definizione, la revisione e la comunicazione degli obiettivi aziendali.
- L'identificazione e l'implementazione di iniziative di miglioramento continuo.
- Il supporto a decisioni basate su dati concreti (*data-driven*) a tutti i livelli dell'organizzazione.

Un approccio strutturato e consapevole alla misurazione della performance è un elemento chiave per il successo e la sostenibilità a lungo termine di qualsiasi organizzazione.

### Prescriptive e Descriptive Knowledge nei Livelli di Controllo

Il controllo organizzativo opera a tre livelli interconnessi: strategico, operativo e tattico. Un controllo efficace richiede sia la **Prescriptive Knowledge** (come le cose dovrebbero essere per raggiungere gli obiettivi) sia la **Descriptive Knowledge** (come le cose sono realmente, basata sull'osservazione).

- **Livello Strategico:**
  - **Prescriptive Knowledge:** Definisce gli obiettivi di alto livello, le specifiche e i target desiderati (es. KPI strategici, Balanced Scorecard). Guida la pianificazione e le priorità.
  - **Descriptive Knowledge:** Deriva dall'osservazione e misurazione dei risultati (es. analisi BI, report aggregati). Descrive la realtà aziendale rispetto agli obiettivi.
- **Livello Operativo:**
  - **Prescriptive Knowledge:** Si manifesta in procedure operative, regole aziendali e allocazione delle risorse (es. modelli di processo *de jure*, manuali).
  - **Descriptive Knowledge:** Ricavata dall'analisi dell'esecuzione dei processi (es. Process Mining, event log). Identifica le differenze tra ciò che è previsto e ciò che accade (*gap analysis*).
- **Livello Tattico:**
  - **Prescriptive Knowledge:** Include regole decisionali operative e configurazioni dei task (es. motori decisionali, workflow configurati).

- **Descriptive Knowledge:** Si ottiene monitorando l'esecuzione dei task e analizzando le performance a breve termine (es. monitor in tempo reale, analisi delle varianti).

**Interconnessione:** I KPI (prescrittivi) definiscono gli obiettivi, mentre le osservazioni e i dati reali (descrittivi) ne misurano il raggiungimento. Il confronto tra i due informa le decisioni a tutti i livelli, guidando l'adattamento strategico e il miglioramento operativo. Un controllo obiettivo richiede l'integrazione di entrambi i tipi di conoscenza.

## Performance Strategica, Operativa e Tattica: Un Quadro Sintetico

La misurazione e il controllo della performance in un'organizzazione si articolano su tre livelli interconnessi: strategico, operativo e tattico.

### 1. Performance Strategica:

- **Focus:** Definizione delle priorità e degli obiettivi a lungo termine dell'organizzazione. Stabilisce la direzione generale e le funzioni di alto livello.
- **Controllo:** Mira a verificare il raggiungimento degli obiettivi generali e l'allineamento organizzativo. Il controllo strategico guida le attività operative e tattiche.
- **Misurazione:** Attraverso **KPI strategici** che valutano il successo nel conseguire gli obiettivi di alto livello. Le misure strategiche dipendono dai risultati operativi e tattici.
- **Supporto:** La **Business Intelligence (BI)** svolge un ruolo cruciale nel monitorare la performance strategica, fornendo feedback sulla validità della strategia e potendo evolvere in una risorsa strategica. Il **KUT (Knowledge Unification Test)** assicura la coerenza tra le conoscenze operative/tattiche e gli obiettivi strategici.

### 2. Performance Operativa:

- **Focus:** Esecuzione concreta e monitoraggio delle procedure aziendali. Riguarda l'allocazione delle risorse, la configurazione dei processi e l'identificazione di malfunzionamenti.
- **Controllo:** Orientato all'**efficienza**, ovvero al raggiungimento degli obiettivi con il minimo dispendio di risorse (tempo, denaro, ecc.).
- **Misurazione:** Tramite **KPI operativi** che quantificano aspetti specifici dei processi come costi, tempi di ciclo e accuratezza.
- **Supporto:** Il **Business Process Management (BPM)** e il **Process Mining** sono centrali per analizzare, comprendere e migliorare l'efficienza dei processi a questo livello.

### 3. Performance Tattica:

- **Focus:** Livello intermedio che connette la strategia all'operatività. Si concentra sulla configurazione e sul monitoraggio dei singoli task all'interno delle procedure.
- **Controllo:** Supporta il **decision making in tempo reale**, rilevando deviazioni e anomalie durante l'esecuzione per consentire aggiustamenti immediati.
- **Misurazione:** Utilizza tecniche di **Process Mining** (come il rilevamento di non conformità e la previsione dei tempi) per monitorare l'esecuzione dei task e delle procedure in tempo reale.

### Relazione Interconnessa tra i Livelli:

- Il **livello strategico** definisce le priorità che guidano le attività operative e tattiche.
- I **KPI operativi e tattici** forniscono un feedback essenziale che alimenta il controllo strategico, influenzando potenzialmente la revisione delle strategie.
- La consapevolezza e la coerenza con gli obiettivi strategici sono cruciali per il successo delle iniziative di **Process Mining** a livello operativo e tattico.

- Sebbene non tutti i processi vengano monitorati nel dettaglio per ragioni di costo ed efficacia, i processi chiave sono soggetti a misurazioni operative e tattiche per garantire l'allineamento strategico e l'efficienza operativa.

### **Performance Indicators: Un Elemento Centrale nella Gestione**

Performance indicators are a central element of management science, sebbene la loro selezione non sia un compito semplice. Esistono numerosi modelli di riferimento come il Supply Chain Operation Reference Model (SCOR), l'American Productivity and Quality Council (APQC) e l'IT Infrastructure Library (ITIL), che possono fornire una guida nella loro scelta. È importante considerare che questi indicatori possono riferirsi a diversi livelli dell'organizzazione: strategico, operativo e tattico.

Il Business Process Management (BPM) si concentra tipicamente sul livello operativo o tattico, con l'obiettivo di eseguire i processi in modo efficiente: raggiungendo gli obiettivi con il minimo sforzo o spesa superflua. Anche il Process Mining (PM) pone una forte enfasi su questi livelli. Tradizionalmente, in questi contesti si considerano tre dimensioni principali per la misurazione della performance: Tempo, Costo e Qualità.

### **Performance Indicators Relativi al Tempo:**

- **Processing Time (PT):** Il tempo impiegato da una singola attività.
- **Waiting Time (WT):** Il tempo che intercorre tra l'esecuzione di diverse attività. Rappresenta il periodo in cui un caso è in attesa che una risorsa diventi disponibile.
- **Lead Time (LT):** Il tempo totale che trascorre tra l'inizio e la fine di un insieme di attività finalizzate al raggiungimento di un obiettivo specifico (istanza o caso di processo). Può essere misurato a livello di singola attività o per l'intero flusso di lavoro.
- **Transition Time (TT):** Il tempo complessivo impiegato per passare da un'attività all'altra all'interno di un'istanza o caso di processo. Corrisponde al Tempo di Attesa (WT).
- **Cycle Time (CT):** Il Lead Time meno il Transition Time (LT - TT). Rappresenta anche la somma di tutti i Tempi di Elaborazione.
- **Cycle Time Efficiency (CTE):** Il rapporto tra il tempo a valore aggiunto (Cycle Time) e il tempo totale (Lead Time), indicando la proporzione di tempo effettivamente dedicata ad attività produttive.
- **Incremental Lead Time (ILT):** L'aumento del Lead Time per ogni attività successiva, utile per identificare i punti di strozzatura all'interno del processo.
- **Incremental Cycle Time (ICT):** L'aumento del Cycle Time per ogni attività successiva, anch'esso utile per localizzare i colli di bottiglia.
- **Average Process Overdue Time:** La media del tempo di superamento delle scadenze previste per i processi.
- **Percentage of Overdue Processes:** La proporzione di processi che non rispettano le tempistiche stabilite.
- **Average Time to Complete Task:** La media del tempo necessario per portare a termine una singola attività.

### **Performance Indicators Relativi al Costo:**

- **Key Cost (KC):** Il costo associato alle attività che apportano valore al prodotto o servizio, ovvero i costi sostenuti durante il tempo di elaborazione per le attività richieste dal cliente.
- **Accessory Cost (AC):** Il costo delle attività che non aggiungono valore, ovvero i costi sostenuti durante il tempo di attesa per le attività non richieste dal cliente.

- **Cost:** La somma dei costi materiali e dei costi relativi alle risorse impiegate nel processo. I costi di esecuzione di un'attività possono essere fissi o variare in base al tipo di risorsa, al suo utilizzo e alla durata dell'attività.
- **Cost of Quality (CoQ):** I costi sostenuti per identificare e risolvere problemi di qualità.
- **Average Consumed Resources (ACR):** La media del costo delle risorse materiali consumate per ogni caso di processo.
- **Stored Resources (SR):** Il valore delle risorse materiali immagazzinate.
- **Storage Support Capacity (SSC):** Una metrica che indica, ad esempio, il numero di giorni di copertura del fabbisogno di risorse materiali garantita dalle scorte.
- **Average Supply Delivery (ASD):** Il numero medio di giorni necessari alla catena di approvvigionamento per consegnare le risorse.
- **Sum of Costs of “Killed” / Stopped Active Processes:** Il valore economico dei processi attivi che sono stati interrotti prima del completamento.

#### **Performance Indicators Relativi alla Qualità:**

- **Product Quality (PdQ):** La qualità del prodotto, misurata attraverso indicatori come il tasso di difettosità e la durabilità.
- **Delivery Quality (DQ):** La qualità della consegna, valutata tramite il tasso di consegna puntuale e la varianza del ciclo di vita.
- **Customer Satisfaction (CS):** La soddisfazione del cliente, misurata attraverso punteggi di feedback dei clienti, il Net Promoter Score, il tasso di fidelizzazione e il tempo di prima risposta.
- **Process Quality (PcQ):** La qualità del processo, che può includere l'engagement dei dipendenti.
- **Defect Rate:** Il tasso di difettosità, ovvero la percentuale di prodotti o servizi che non soddisfano gli standard di qualità.
- **Employee Engagement:** Il livello di coinvolgimento dei dipendenti, che indica la loro motivazione e partecipazione al processo.
- **Number of Process Errors vs Number of Human Errors:** Il numero di errori di processo confrontato con il numero di errori umani, per distinguere le problematiche sistemiche da quelle individuali.

#### **Altri Performance Indicators:**

- **Resource Utilisation (RU):** L'utilizzo delle risorse, ovvero il rapporto tra il tempo effettivo di impiego di una risorsa e il tempo totale disponibile per tale risorsa. È importante notare che un aumento dell'utilizzo delle risorse può portare a maggiori tempi di attesa a causa della formazione di code quando le risorse sono costantemente occupate.
- **Revenue (Rv):** Il ricavo generato da un'istanza di processo.
- **Return (Rt):** Il ritorno economico, calcolato come la differenza tra il ricavo e la somma del Costo Chiave e del Costo Accessorio ( $Rv - (KC + AC)$ ).
- **Work-in-process (WIP):** Il lavoro in corso, ovvero il numero di istanze di processo attive in un dato momento.
- **Case Arrival Rate (CAR):** Il tasso di arrivo dei casi, ovvero il numero medio di nuovi casi creati per unità di tempo.
- **Inter-arrival Time (IAT):** Il tempo inter-arrivo, ovvero il tempo medio che intercorre tra l'arrivo di due nuovi casi.
- **Average Worked Cases (AWC):** La media dei casi lavorati, ovvero il numero medio di casi completati per unità di tempo.
- **Volume of Tasks per Staff:** Il volume di compiti per personale, che indica il carico di lavoro medio per dipendente.



- **Percentage of Processes where the actual number assigned resources is less than planned number of assigned resources:** La percentuale di processi in cui il numero effettivo di risorse assegnate è inferiore al numero pianificato.
- **Time allocated for activity type: administration, management, training:** Il tempo allocato per tipologia di attività, come amministrazione, gestione e formazione.

La definizione di un indicatore di performance implica la misurazione di una o più dimensioni quantificabili degli esiti dell'organizzazione, e i valori che l'indicatore può assumere sono definiti come livelli. Un indicatore può essere collegato a un obiettivo che specifica il livello desiderato o il miglioramento atteso. L'analisi di questi indicatori, spesso supportata da tecniche di Process Mining basate sui dati degli event log, è fondamentale per valutare le prestazioni aziendali rispetto agli obiettivi e per identificare opportunità di miglioramento continuo. Il Process Performance Management si concentra sul monitoraggio di questi indicatori per ottimizzare l'efficienza e la qualità dei processi.

## DATA MODEL

Per misurare le performance, è fondamentale disporre di un **data model** capace di indirizzare tutte le dimensioni rilevanti, con particolare enfasi sul tempo. Come hai giustamente sottolineato, il tempo è una dimensione cruciale per valutare le performance delle organizzazioni.

Un modello dati ideale per l'analisi delle performance dei processi dovrebbe includere informazioni dettagliate come:

- **Timestamp** di inizio e fine per ogni attività, che consentono di calcolare il **Processing Time (PT)**, il **Waiting Time (WT)** tra le attività e, di conseguenza, il **Lead Time (LT)**, che è la somma del tempo impiegato da tutte le attività e di tutti i tempi tra di esse. La presenza di timestamp di inizio e completamento per ogni evento è necessaria se siamo interessati a calcolare la durata delle attività.
- La **sequenza delle attività o dei task** eseguiti per raggiungere uno specifico obiettivo. Questa sequenza definisce una **Case** o **Process Instance** o **Process Execution**. Una **Variant** è definita come una collezione di casi che seguono la stessa sequenza di attività. L'ordine degli eventi all'interno di un caso è cruciale per scoprire le dipendenze causali nei modelli di processo.
- Informazioni sulle **risorse** coinvolte nell'esecuzione delle attività.
- Eventuali **costi** associati alle attività.
- Altri attributi rilevanti per l'analisi, come l'**esito dell'attività**.

Come evidenziato, un problema tipico si presenta quando nei nostri dataset è disponibile solo il **timestamp di inizio**. In questo scenario, non siamo in grado di calcolare il tempo di processing e il tempo di attesa, e di conseguenza l'efficienza del ciclo di vita (**Cycle Time Efficiency**) diventa problematica. Il Cycle Time Efficiency (CTE) è il rapporto tra il tempo a valore aggiunto e il tempo non a valore aggiunto.

Le opzioni disponibili in caso di modello dati con singolo timestamp possono includere:

- Calcolare il **Processing Time** basandosi sul timestamp dell'attività successiva. Tuttavia, come giustamente noti, questo approccio potrebbe funzionare solo con processi molto regolari e nella maggior parte dei casi non ci permetterà di conoscere il **Waiting Time**. Inoltre, se le attività sono eseguite in parallelo, l'attività successiva potrebbe iniziare prima che una precedentemente iniziata sia terminata.
- Calcolare il **Lead Time** per caso, prendendo il primo e l'ultimo evento di un caso. Questo non ci permette di calcolare il **Cycle Time**, ma consente di studiare la relazione tra il **Lead**

**Time** del caso e la sua complessità. Il Lead Time (LT) è il tempo totale dalla creazione del caso al completamento. Il Cycle Time (CT) è il Lead Time meno il Transition Time, o anche la somma di tutti i Tempi di Elaborazione (PT).

- Calcolare l'**Incremental Lead Time (ILT)**. Questo non consente di calcolare il **Cycle Time**, ma permette di identificare i colli di bottiglia. L'Incremental Lead Time è l'incremento del Lead Time per ogni attività ed è utile per identificare i colli di bottiglia.

È importante sottolineare che, per un'analisi più approfondita delle performance, come l'identificazione di colli di bottiglia, la misurazione dei livelli di servizio e il monitoraggio dell'utilizzo delle risorse, la presenza di timestamp di inizio e fine (o almeno di completamento) per gli eventi è cruciale. Un log degli eventi ideale dovrebbe registrare l'inizio di un'attività con un evento di inizio, il completamento con un evento di fine ed, eventualmente, anche eventi di interruzione o ripresa. Per ogni evento, dovrebbe essere noto il tempo di occorrenza tramite un timestamp.

Come menzionato, gli event log dovrebbero essere trattati come "cittadini di prima classe" e i sistemi informativi dovrebbero fornire i mezzi per registrare (semi)automaticamente dati di log orientati al processo. La qualità dei dati, inclusa la precisione e la granularità dei timestamp, è fondamentale per ottenere risultati affidabili nell'analisi dei processi. L'affidabilità nella registrazione dei timestamp è importante.

In sintesi, sebbene sia possibile effettuare alcune analisi con un modello dati con solo timestamp di inizio, la mancanza di timestamp di fine limita significativamente la capacità di misurare accuratamente metriche chiave come il **Processing Time**, il **Waiting Time** e l'**Cycle Time Efficiency**, ostacolando una comprensione completa delle performance del processo e l'identificazione efficace di aree di miglioramento come i colli di bottiglia. Per un'analisi approfondita delle performance, sono necessari log degli eventi che includano informazioni temporali dettagliate sull'inizio e la fine delle attività.

## Lezione 4

### Process Mining: Introduzione

- **Definizione:** Disciplina emergente tra data science e process science.
- **Obiettivo:** Trasformare i dati degli eventi in insight e azioni per migliorare i processi operativi.
- **Focus:** Analisi dei processi basata sui log di eventi.
- **Analisi "As-Is":** Scoprire, monitorare e migliorare processi reali.
- **Collega:** Analisi tradizionale basata sui modelli (BPM) e tecniche data-centriche (machine learning, data mining).
- **Non Semplice Unione:** Costruisce su approcci esistenti ma aggiunge una prospettiva unica.
- **Prospettiva di Processo:** Aggiunge la dimensione del processo al machine learning e al data mining (a differenza di tecniche "process-agnostic").
- **Applicazione Semplificata:** Facilita l'applicazione di tecniche di data mining ai dati di eventi.
- **Insight Precisi:** Offre insight più accurati rispetto agli strumenti di Business Intelligence (BI) (oltre a dashboard e reporting).
- **Confronto:** Analizza la discrepanza tra il processo osservato (log di eventi) e il processo atteso/modellato.

### Process Mining: Nozioni Preliminari

#### Dati Fondamentali: I Log di Eventi (Event Logs)

- **Essenziali:** Il Process Mining richiede log di eventi adeguati.
- **Definizione:** Collezione di eventi registrati da un sistema informativo.
- **Contenuto di Ogni Evento:**
  - Attività/Task eseguito.
  - Momento specifico.
  - Caso particolare.
- **Origine Dati:** Non creano nuovi dati, ma derivano da altre basi di dati e audit trail.
- **Attributi Aggiuntivi degli Eventi:**
  - Timestamp dell'evento.
  - Risorsa (persona o dispositivo).
  - Informazioni sui dati (risultato, costo).
  - Tipo di transazione (start, complete, suspend).
  - Informazioni relative al cliente o alla posizione.
- **Estrazione, Integrazione e Trasformazione:** Fondamentali e spesso impegnative.
- **Fonti Dati:** File (XES, MXML, Excel, CSV), database, adapter specifici.
- **Standard Supportati:** XES (eXtensible Event Stream), MXML (Mining eXtensible Markup Language).
- **Importanza di Scoping:** Selezione dei dati rilevanti.
- **Gestione della Qualità:** Affrontare problemi di dati e rumore nei log.
- **Necessità di Log di Alta Qualità:** Essenziale per risultati di valore.
- **Importanza della Provenance:** Registrazione sistematica e affidabile degli eventi.

#### Casi, Trace e Varianti

- **Caso (Case):** Collezione di eventi relativi alla stessa esecuzione del processo (richiede un ID di istanza).

- **Trace:** Sequenza di eventi per un caso (ordinati tramite timestamps).
- **Variante:** Collezione di casi che seguono la stessa trace.
- **Distribuzione di Pareto:** Numero limitato di trace frequenti, molte trace poco frequenti.
- **Profilazione delle Varianti:** Punto di partenza per comprendere il processo.
- **Analisi delle Varianti:** Mira a spiegare le differenze tra i casi eseguiti.

## Modelli di Processo (Process Models)

- **Definizione:** Collezione di specifiche logiche che prescrivono sequenza, sincronizzazione, pre/post-condizioni per raggiungere un obiettivo.
- **Distinzione:** Rappresentazione (linguaggio di modellazione) e comportamento (trace consentite).
- **Linguaggi di Modellazione:**
  - Reti di Petri (Petri nets) e WF-nets.
  - BPMN (Business Process Model and Notation).
  - Diagrammi di attività UML.
  - EPCs (Event-driven Process Chains).
  - Process Trees.
  - Automata.
  - Modelli a basso livello (transition systems, Markov models).
- **Comportamento del Modello:** Insieme delle trace consentite (anche la distribuzione delle trace è importante).
- **Origine dei Modelli:** "Hand-made" (esperti) o scoperti automaticamente dai log di eventi.
- **Molteplici Viste:** Non esiste necessariamente un unico modello "corretto".

## Tipi Principali di Process Mining

### 1. Process Discovery (Scoperta del Processo):

- **Input:** Log di eventi.
- **Output:** Modello di processo (senza informazioni a priori).
- **Obiettivo:** Derivare il modello sottostante che ha generato i log.
- **Tipo di Apprendimento:** Non supervisionato.
- **Tecniche:**
  - Alpha-algorithm ( $\alpha$ -algorithm): Log di eventi  $\rightarrow$  Rete di Petri.
  - Heuristic Miner: Più resiliente al rumore (usato anche in Change Mining).
  - Genetic Miner: Basato su algoritmi evolutivi (per log rumorosi).
  - Inductive Miner: Divide-et-impera  $\rightarrow$  Process Trees (correttezza formale).
  - Fuzzy Mining: Semplificare processi complessi.
  - Tecniche basate sulle regioni o su due fasi.
  - Change Mining: Heuristic Miner su "change logs"  $\rightarrow$  grafi dei cambiamenti.

### 2. Conformance Checking (Verifica di Conformità):

- **Input:** Modello di processo esistente e log di eventi dello stesso processo.
- **Obiettivo:** Verificare quanto bene la realtà (log) si conforma al modello e viceversa.
- **Tipo di Apprendimento:** Supervisionato.
- **Importanza:** Controllo della conformità e auditing.
- **Funzionalità:** Rilevare, localizzare e spiegare le deviazioni.
- **Tipi di Deviazioni:** Attività Insert (log ma non modello), Skip (modello ma non log), Early/Late, problemi di sincronizzazione, parallelizzazione, iterazioni, pre/post-condizioni, risorse.
- **Tecniche:** Replay basato su token, allineamenti (alignments).

- **Metriche:** Conformità (fitness, precisione), controllo di regole specifiche/business rules.

### 3. Enhancement (Miglioramento):

- **Input:** Modello di processo esistente e log di eventi.
- **Obiettivo:** Estendere o migliorare il modello usando informazioni dal log.
- **Azioni:** Riparazione (modificare il modello), estensione (aggiungere prospettive: prestazioni, risorse, regole).

### Altre Prospettive e Analisi

- **Prospettiva Organizzativa (Organizational Mining):** Analisi relazioni tra risorse e strutture organizzative.
- **Prospettiva dei Dati:** Analisi dell'influenza dei valori dei dati sul processo (regole decisionali).
- **Prospettiva Temporale e delle Prestazioni:** Analisi tempi di attesa, throughput, colli di bottiglia, utilizzo risorse, livelli di servizio (tramite timestamps).

### Supporto Operativo (Online Process Mining)

- **Applicazione in Tempo Reale:** Supporto ai processi operativi in esecuzione ("pre mortem").
- **Attività:**
  - **Detect:** Rilevare deviazioni/violazioni in tempo reale.
  - **Predict:** Previsioni sul futuro del caso (tempo rimanente, probabilità di deviazione).
  - **Recommend:** Raccomandare azioni appropriate.
- **Funzionalità Simili a GPS:** Navigazione processi, proiezione informazioni dinamiche, previsione tempi.

### Strumenti

- **Necessità:** Strumenti dedicati (software generico di data mining/BI non ottimizzato).
- **Open-Source:**
  - ProM (Process Mining framework): Piattaforma principale (ProM 5.2, ProM 6).
  - RapidProM: Estensione di RapidMiner con plug-in ProM.
  - PM4PY.
  - PMTK.
  - BPMN.io.
  - cortado.
  - RuM.
- **Creazione di Modelli Specifici:**
  - PNML Framework (Petri Nets).
  - APO (Petri Nets).
  - BPMN.IO (BPMN).
  - BPMNDiffViz (BPMN).
  - Cortado (Interactive process discovery).
  - RuM (Declarative modeling).
- **Generazione di Log Sintetici:**
  - PLG2.
  - BIMP.
  - L-Sim.
  - Loggenerator (ProM plugin).
  - PURPLE.

- **Strumenti Commerciali:** Offrono funzionalità di Process Mining (variabile supporto per modelli e attività, scalabilità generalmente buona).

## Metodologie e Tipi di Processi

- **Metodologie:** Guidare i progetti (es. L\* life-cycle model: pianificazione, estrazione, creazione modelli, integrazione, supporto operativo).
- **Guida dei Progetti:** Dati, domande specifiche, obiettivi (KPIs).
- **Tipi di Processi (Continuum):**
  - "Lasagna processes": Strutturati, ripetibili.
  - "Spaghetti processes": Meno strutturati, alta variabilità (analisi più complessa).
- **Tecniche per Processi "Spaghetti":** Filtraggio log, clustering casi, fuzzy mining.

## Sfide Aperte

- Disponibilità e qualità dei dati.
- Rumore e incompletezza dei log.
- Gestione del "concept drift" (cambiamenti nel tempo).
- Scalabilità per log molto grandi ("Big Event Data").
- Complessità della scoperta di modelli (bilanciare fitness e precisione).
- Integrazione con altre tecniche di analisi (data mining).
- Questioni etiche (privacy, sicurezza).

## Process Mining Use Cases

- **Ottimizzazione del Processo (Process Optimization):**
  - Analisi rapida e accurata (metriche, colli di bottiglia, costi).
  - Innescare azioni di miglioramento (redesign, aggiustamenti, interventi).
  - Basato su KPIs (tempo, costi, qualità).
  - Diagnosi dei problemi basata sui fatti.
- **Scoperta del Processo per l'Automazione (Process Discovery for Automation):**
  - Esaminare i processi per un'automazione efficiente.
  - Scoprire i processi effettivi ("as-is" da log).
  - Comprendere il comportamento reale prima dell'automazione.
  - Fornire insight sull'utilizzo dei sistemi.
- **Verifica di Conformità (Conformance Validation):**
  - Verificare la conformità dei processi effettivi alle specifiche.
  - Confronto modello esistente vs. log di eventi.
  - Rilevare, localizzare e spiegare le deviazioni.
  - Cruciale per auditing e compliance (es. regole specifiche).
  - Supporta la verifica di regole basate sui dati (decision mining/DPA).
- **Simulazione del Processo (Process Simulation / Previsione):**
  - Fare previsioni future basate sui dati dei log.
  - Prevedere ritardi, tempi rimanenti, probabilità di non conformità.
  - Supporto operativo ("pre mortem").
  - Fornire informazioni per stakeholder e decisioni.
  - Formulare raccomandazioni.
  - Supportato da strumenti dedicati.

## Process Mining Software and Libraries

- **General purpose:**
  - PM4PY
  - ProM
  - BupaR
- **Declarative modeling:**
  - RuM
- **Create BPMN:**
  - BPMN.IO
  - BPMNDiffViz
- **Interactive process discovery:**
  - Cortado
- **Create Petri Nets:**
  - PNML Framework
  - APO
- **Synthetic event logs:**
  - PLG2
  - BIMP
  - L-Sim
  - Loggenerator (ProM plugin)
  - PURPLE

## Event Logs

- **Definizione:** Collezione di eventi registrati da un sistema informativo (attività, momento, caso).
- **Struttura Dati Fondamentale:** Per l'analisi orientata ai processi (Process Mining).
- **Varietà di Forme:** Diverse implementazioni di logging nei sistemi.
- **Origine Dati:** Sistemi informativi, audit trail, log di transazione, database, data warehouse.
- **Dati Nascosti o Dispersi:** Spesso sottoprodotto per il debugging o distribuiti su più tabelle.
- **Perdita di Dati:** Alcuni sistemi sovrascrivono i vecchi valori durante gli aggiornamenti.
- **Formati Standard:** MXML, XES, CSV.
- **eXtensible Event Stream (XES):** Standard IEEE per log ed stream di eventi (basato su XML), successore di MXML (più estendibile, senza attributi obbligatori predefiniti), le estensioni forniscono semantica.
- **Campi Obbligatori ("Minimo Indispensabile"):**
  - Caso ID (identificatore dell'istanza).
  - Nome Attività (Activity name).
  - Ordinamento degli eventi all'interno del caso (tipicamente tramite Timestamp).
- **Altri Campi di Interesse:** Per KPI e statistiche (risorsa, timestamp, elementi dati).
- **Informazioni Temporalistiche Dettagliate:** Richiedono Timestamp di Avvio e di Completamento per il tempo di elaborazione.
- **Tipo di Transazione:** Specifica la fase del ciclo di vita (start, complete, schedule).
- **Molteplici Dimensioni Catturate:** Flusso di controllo (Control-Flow), dati, organizzazione, caso (Case), risorse (Resource), tempo (Time).

## Misurazione delle Performance (Performance Measurement)

- **Definizione:** Attività essenziale per valutare il raggiungimento di obiettivi e standard.
- **Scopo:** Processo di rendicontazione sulla performance di un'organizzazione.
- **Funzione di Controllo del Management:** Ortopogonale a pianificazione, organizzazione, leadership, staffing.
- **Impatto:** Valutazione e processo decisionale a livello strategico, operativo e tattico.
- **Business Intelligence (BI):**
  - Parte da un obiettivo (spesso KPI).
  - **KPI (Indicatori Chiave di Performance):**
    - Essenziali per il management.
    - Collegano attività agli obiettivi.
    - Definiscono quantità misurabile (performance di processo o business).
    - Tipi:
      - Quantitativi (numeri).
      - Pratici (interfaccia con i processi).
      - Direzionali (indicano miglioramento).
      - Azionabili (controllano effetti del cambiamento).
      - Finanziari.
    - Identificazione basata su:
      - Processo di business predefinito.
      - Requisiti del processo.
      - Misurazione dei risultati rispetto agli obiettivi.
    - Implicano la misurazione di una o più dimensioni misurabili.
- **Dimensioni della Performance di un Processo:**
  - **Tempo:** Tempo di attraversamento (throughput time), tempo di servizio (service time), tempo di attesa (waiting time), tempo di sincronizzazione (synchronization time).
  - **Costo:** Costi di esecuzione di attività, risorse, ecc.
  - **Qualità:** Compliance, soddisfazione cliente, numero di difetti, qualità prodotto/ consegna/processo, costi della qualità, work-in-process (WIP).
  - **Flessibilità:** Difficile da quantificare.
- **Process Mining e Performance Management:**
  - Process Mining: Livello operativo o tattico (efficienza dei processi).
  - Process Performance Management: Definizione, supervisione e valutazione di KPI specifici per l'esecuzione dei processi.
- **Criticità della Misurazione Aggregata:**
  - Misurare a livello di processo (generale) è spesso fuorviante.
  - I valori medi non sono affidabili (distribuzione di Pareto delle varianti).
  - Risultati influenzati da dati incompleti, errati o rari (rumore).
  - Necessità di termini di confronto.

## Analisi delle Varianti (Process Variant Analysis)

- **Definizione:** Tecniche per analizzare log di eventi e spiegare le differenze tra i casi eseguiti.
- **Scopo:** Comprendere le differenze per standardizzare o migliorare un processo.
- **Variante:** Collezione di casi con la stessa Traccia.
- **Distribuzione delle Varianti:** Tipicamente di Pareto (poche frequenti, molte infrequenti).
- **Profilazione delle Varianti:** Punto di partenza per la comprensione.



- **Variante Rilassata:** Casi conformi a uno specifico pattern.
- **Affronta le Limitazioni:** Supera la misurazione aggregata e i valori medi non affidabili.
- **Livello di Analisi:** Da processo generale a varianti o segmenti di casi.
- **Valutazione della Performance di una Variante/Segmento:** Confronto con un livello di riferimento.
- **Livelli di Riferimento:**
  - Obiettivo definito dall'organizzazione (confronto booleano o tolleranza).
  - Livelli/frequenze di altre varianti (analisi della significatività della differenza).
- **Sintesi:** L'analisi a livello aggregato può essere fuorviante. L'analisi delle varianti fornisce il dettaglio necessario per comprendere le differenze di performance e guidare il miglioramento (confronto con obiettivi o altre varianti/segmenti). La qualità dei dati influenza l'accuratezza dell'analisi.

## FILTERING NOISE

- **Definizione:** Informazioni errate, imprecise o incomplete nel log di eventi.
- **Impatto:** Compromette la qualità dell'analisi di Process Mining.
- **Non Solo Errori di Logging:** Anche comportamenti rari o atipici (outliers).
- **Difficoltà di Distinzione:** Algoritmi faticano a distinguere errori da eventi eccezionali (richiede intervento umano).
- **Esempi di Rumore:**
  - Errori di registrazione eventi/casi.
  - Attività non rilevanti o con durata nulla.
  - Timestamps errati o incompleti (troppo granulari, diversi dal reale).
  - Ordine scorretto degli eventi (causato da timestamp mancanti/imprecisi).
  - Casi incompleti (terminano in WIP, durata oltre l'osservazione).
  - Eventi incompleti (mancano timestamp o attributi).
  - Casi che iniziano con attività inappropriate.
  - Incompletezza generale (difficoltà a scoprire la struttura reale).
- **Principio 80/20:** Obiettivo di descrivere l'80% del comportamento, tralasciando il 20% anomalo.
- **Tecniche Avanzate di Discovery:** Heuristic Mining, Fuzzy Mining, Genetic Mining, Inductive Mining (progettate per gestire rumore e incompletezza).
- **Scopo del Filtraggio:** Migliorare la comprensione del processo.
- **Necessità di Intervento Umano:** Pre-elaborazione e post-processing del log.

## FILTERING IRRELEVANT DATA

- **Definizione:** Informazioni corrette ma non pertinenti per l'analisi specifica.
- **Correlazione con lo Scoping:** Definizione del perimetro dell'analisi.
- **Guida dell'Estrazione:** Basata sulle domande da rispondere.
- **Esempi di Dati Irrilevanti:**
  - Intervalli temporali non di interesse.
  - Casi gestiti da dipartimenti/risorse non rilevanti.
  - Casi con attività/percorsi fuori dallo scope.
  - Attività obsolete o marginali.
- **Scopo del Filtraggio:** Semplificare la vista e migliorare la comprensione.
- **Azioni di Filtraggio:**
  - Includere solo le varianti più frequenti.
  - Includere solo le attività centrali.
  - Rimuovere casi lenti o eccezionali.
  - Focalizzarsi su specifici segmenti del processo.
- **Processo Iterativo:** Analisi esplorativa per affinare i filtri.

## Filtering for Summarisation

- **Problema:** Log con troppa informazione (difficoltà di comprensione).
- **Scopo:** Semplificare la rappresentazione mantenendo le informazioni rilevanti.
- **Obiettivi:**
  - Vista semplificata del comportamento.
  - Evitare modelli troppo complessi.
  - Migliorare la leggibilità dei diagrammi.
- **Strategie Comuni:**
  - **Filtrare le varianti più frequenti:** Basato sulla distribuzione di Pareto (focalizzarsi sull'80% dei casi).
  - **Rilassare la nozione di variante:** Casi con stessa attività iniziale/finale, presenza di una certa attività, cycle time entro un intervallo (forma estesa di analisi delle varianti per identificare segmenti significativi).
  - **Filtrare le attività più frequenti:** Rimuovere attività con bassa frequenza (es. < 5% dei casi). (Supportato da Heuristic Miner per l'astrazione).
  - **Filtrare i percorsi più frequenti:** Rimuovere percorsi marginali per modelli più comprensibili (collegato al Rasoio di Occam).

## Conclusione

- **Filtraggio dei Log di Eventi:** Passaggio cruciale e spesso sottovalutato nel Process Mining.
- **Tre Tipi Principali:**
  - **Filtering Noise:** Rimuove errori e comportamenti atipici.
  - **Filtering Irrelevant Data:** Elimina ciò che non è utile per gli obiettivi dell'analisi.
  - **Filtering for Summarisation:** Semplifica la vista per la comprensione.
- **Benefici del Filtraggio Corretto:** Migliora la qualità dei modelli scoperti e consente analisi più accurate, efficaci e orientate alle decisioni.

## Typical filters in process mining

Filter category	Description
Start Activities	Filter by the initial activity <i>initial</i> in cases
End Activities	Filter by the concluded activity <i>final</i> in cases
Directly-Follows	Select cases based on specific pairs of consec. activities
Prefixes	Filter by initial sequences of activities in cases
Suffixes	Filter by final sequences of activities in cases
Case Size	Filter by the length or number of activities in cases
Time	Filter cases or events within a specified period
Throughput	Filter based how long cases take to complete
Attributes	Filter by specific values of categorical attributes
String Attributes	Filter by specific values or ranges of numerical attributes
Numeric Attributes	Filter by specific values or ranges of attributes

# Perché i Valori Medi a Livello di Processo Sono Spesso Fuorvianti

**Premessa:** La misurazione delle performance è fondamentale tramite indicatori (KPIs), e il Process Mining estrae metriche dai log eventi. Tuttavia, **i valori medi aggregati a livello di processo sono intrinsecamente inaffidabili e fuorvianti.**

**Motivazioni:** I processi reali presentano una **variabilità significativa**, che i valori medi non catturano.

## 1. Variabilità e Complessità del Processo:

- I log eventi rivelano diverse **varianti** (sequenze di attività uniche) e comportamenti non uniformi.
- Un valore medio (es. tempo di ciclo medio) è influenzato da tutte le varianti, incluse quelle rare.
- **Non indica:** quali varianti sono più veloci/lente o dove si verificano i ritardi specifici.

## 2. Rumore e Comportamento Infrequente:

- I log contengono quasi sempre **rumore** (comportamenti rari e non tipici/outlier).
- Gli outlier (es. casi con tempi di ciclo estremi dovuti a errori) possono **distorcere significativamente i valori medi**, fornendo un'immagine errata della performance tipica.

## 3. Fattori Contestuali e Stato Non Stazionario:

- I processi sono influenzati da fattori esterni (orari, giorni, stagioni) e raramente sono in uno stato stabile.
- Un valore medio su un lungo periodo **maschera pattern temporali importanti** (es. rallentamenti in specifici momenti) o l'impatto di eventi isolati.
- **Soluzione preferibile:** considerare intervalli di confidenza o analizzare i dati per segmenti temporali.

## 4. Problemi di Qualità dei Dati:

- Errori di inserimento, dati mancanti o attributi errati sono comuni nei log reali.
- Valori medi calcolati su dati di bassa qualità sono **intrinsecamente inaffidabili.**

## 5. Necessità di Segmentazione (Variant Analysis):

- L'analisi aggregata a livello di processo è meno informativa dell'analisi granulare.
- L'**Analisi delle Varianti** e il **filtering** permettono di suddividere i casi in segmenti omogenei (per attività di inizio/fine, inclusione di specifiche attività, intervalli di tempo di ciclo, risorse coinvolte).
- **Misurare le performance su questi segmenti** (o "varianti rilassate") fornisce insight molto più affidabili e confrontabili.
- Permette di identificare colli di bottiglia specifici per determinate varianti e comprendere le cause della variabilità.

## In Sintesi:

- La misurazione delle performance è cruciale, ma i **valori medi aggregati sull'intero processo sono spesso fuorvianti** a causa di variabilità, rumore e molteplicità di comportamenti.

- Il Process Mining supera questo limite fornendo strumenti per **analizzare i dati a diversi livelli di astrazione** e per **segmentare i log eventi**.
- L'**analisi delle varianti** e il **confronto tra segmenti** offrono una comprensione delle performance più dettagliata e affidabile.

## Distribution of Variants

### Variant Analysis: Comprendere le Differenze nell'Esecuzione dei Processi

- **Definizione:** Insieme di tecniche per analizzare i log eventi e spiegare le differenze tra i casi eseguiti.
- **Obiettivi:** Comprendere le differenze, supportare decisioni informate su standardizzazione/miglioramento, rispondere a domande su frequenza, similarità e performance delle varianti.
- **Concetto di Variante:**
  - Collezione di casi che seguono la stessa **trace** (sequenza di attività).
  - Definizione **rilassata:** Insieme di casi che soddisfano specifici vincoli (stessa attività iniziale/finale, inclusione di attività, soglie temporali, risorse).
  - Analisi su **qualsiasi segmento** del Log Eventi.
- **Esempio:** Illustrazione di come diverse sequenze di attività definiscono varianti.

### Distribuzione delle Varianti: La Legge di Potenza

- La frequenza delle varianti segue tipicamente una **legge di potenza**.
- **Principio di Pareto (80/20):** Poche varianti rappresentano la maggior parte dei casi.
- **Comportamento Principale (Mainstream):** Varianti frequenti.
- **Comportamento Non-Mainstream:** Varianti rare, ma eterogenee e spesso causa di problemi di performance/compliance.
- **Visualizzazione:** Grafico illustrativo della distribuzione (poche varianti con alta frequenza, molte con bassa frequenza).

### Funzioni di Densità di Probabilità (PDF), Cumulativa (CDF) e Complementare (CCDF)

- **Probability Density Function (PDF):** Probabilità che una variabile casuale (frequenza di una variante) assuma un valore specifico.
- **Cumulative Density Function (CDF):** Probabilità che una variabile casuale (frequenza di una variante) sia minore o uguale a un certo valore.
- **Complementary Cumulative Density Function (CCDF):** Probabilità che una variabile casuale (frequenza di una variante) sia maggiore di un certo valore.
- **Applicazioni:** Comprendere la forma della distribuzione, quantificare la rarità/dominanza delle varianti.

### Analisi della Distribuzione delle Varianti: Legge di Potenza e Identificazione

- **Concetto Chiave:** Poche varianti spiegano la maggior parte dei casi.
- **Modellazione:**  $P(x) \propto x^{-\alpha}$  ( $\alpha$ : esponente,  $x_{min}$ : cutoff).
- **Identificazione dei Parametri ( $\alpha$  e  $x_{min}$ ):**

1. **Estrarre le frequenze delle varianti** dal log eventi. Questo si traduce in una lista o array dove ogni elemento rappresenta la frequenza con cui una specifica variante è stata osservata.
  2. **Adattare la legge di potenza ai dati utilizzando librerie statistiche** come `powerlaw` in Python. La funzione `powerlaw.Fit()` esegue questo adattamento.
 

```
python import powerlaw
counts = # Frequenze delle varianti (es. [100, 50, 30, 10, 5, 2, 1, 1])
fit = powerlaw.Fit(counts)
print(f"Alpha: {fit.power_law.alpha:.2f}, x_min: {fit.power_law.xmin}")
```
  3. **Stima dei parametri:** La libreria `powerlaw` utilizza metodi di massima verosimiglianza (Maximum Likelihood Estimation - MLE) per stimare i valori di  $\alpha$  (l'esponente della legge di potenza) e  $x_{min}$  (il cutoff minimo a cui la legge di potenza inizia a descrivere bene i dati). L'output `fit.power_law.alpha` conterrà il valore stimato di  $\alpha$ , e `fit.power_law.xmin` conterrà il valore stimato di  $x_{min}$ .
- **Parametri di Interesse:**
    1. **Esponente ( $\alpha$ ):** Tasso di diminuzione della frequenza delle varianti. Un valore più alto indica una diminuzione più rapida.
    2. **Separazione del Segnale ( $x_{min}$ ):** Soglia di frequenza per l'applicazione affidabile della legge di potenza. Le varianti con frequenza inferiore a  $x_{min}$  potrebbero seguire distribuzioni diverse.
    3. **Copertura Varianti Principali:** Spesso il top 1-2% delle varianti copre circa l'80% dei casi.
    4. **Segnalazione (Esempio):** Con  $\alpha \approx 2.1$ , la frequenza della prossima variante più frequente dovrebbe essere circa  $prev^{-2.1}$ . Con  $x_{min} = 28$ , la legge di potenza si applica prevalentemente a frequenze  $\geq 28$ .
  - **Quando la Legge di Potenza Fallisce ( $x < x_{min}$ ):** Utilizzare modelli alternativi per varianti rare (esponenziale, Poisson, lognormale, binomiale negativa). La libreria `powerlaw` può anche aiutare a confrontare l'adattamento della legge di potenza con queste alternative.
  - **Test Statistici:**
    1. **Kolmogorov-Smirnov (KS):** Misura la distanza  $D$  tra la CDF empirica e quella teorica della legge di potenza (basso  $D$  = buon fit).
    2. **Likelihood Ratio:** Confronta la verosimiglianza del modello a legge di potenza con modelli alternativi ( $R > 0$  favorisce la legge di potenza).
  - **Risultati Azionabili:**
    1. **Varianti Principali:** Ottimizzare i percorsi più frequenti per massimizzare l'impatto.
    2. **Varianti Rare:** Investigare se sono errori (necessitano correzione) o innovazioni (potenziale per miglioramenti generalizzati).

## Analisi Comparativa delle Performance e del Controllo di Flusso

### Termini di Comparazione

- La **misurazione delle performance** è essenziale per valutare il raggiungimento di obiettivi e standard.
- Misurare le performance a **livello di processo aggregato** è spesso **fuorviante** a causa della variabilità.

- Sono necessari **termini di comparazione** per un'analisi significativa.

## Filtering per la Comparazione

- Per valutare la performance di un log eventi, è necessario un **termine di comparazione**.
- Possibili termini di comparazione:
  - **Confrontare varianti:** Analizzare le performance tra diverse sequenze di attività.
    - Si può **rilassare la nozione di variante** considerando segmenti del log eventi basati su:
      - Attività di inizio e fine comuni.
      - Inclusione di attività specifiche.
  - **Confrontare timeframe:** Analizzare l'evoluzione delle performance nel tempo.
  - **Confrontare dimensioni organizzative:** Analizzare le performance tra dipartimenti, paesi, ecc.
- Per garantire l'affidabilità, è necessaria **analisi statistica** per valutare la significatività delle differenze osservate.

## Performance Analysis: Indicatori

- Per eseguire l'analisi delle performance su varianti o segmenti, è necessario identificare un **insieme di indicatori**.
- Un **indicatore** implica la misurazione di una o più dimensioni misurabili degli esiti dell'organizzazione.
- Con indicatori multidimensionali, è necessaria una **funzione di aggregazione** per ottenere un singolo valore.
  - Es. **percentuale di casi in ritardo:**  $\text{casi totali} \times \text{casi in ritardo} \times 100$
- I valori che l'indicatore può assumere sono spesso definiti **livelli**.
- L'indicatore può essere collegato a un **obiettivo** che specifica il livello desiderato o il miglioramento atteso.

## Performance Analysis: Indicatori Specifici

- **Tempo:**
  - **Processing time (PT):** Tempo impiegato da un'attività.
  - **Waiting time (WT):** Tempo tra le attività.
  - **Cycle time (CT):**  $PT + WT$ , tempo tra l'inizio e la fine di un insieme di attività (istanza o caso).
  - **Cycle time efficiency (CTE):**  $PT / CT$ , rapporto tra tempo a valore aggiunto e tempo non a valore aggiunto.

## Performance Analysis: Prospettive

- Possono essere considerate **molteplici prospettive** per l'analisi delle performance:
  - **Prospettiva del caso:** Efficienza media del ciclo di vita dei casi in una variante.
  - **Prospettiva della risorsa:** Carico di lavoro di una risorsa in una variante.
  - **Prospettiva del task:** Task con tempi di attesa maggiori per variante.

## Control-Flow Analysis: Pattern

- Per eseguire l'analisi del controllo di flusso su varianti o segmenti, è necessario identificare un **insieme di pattern**.
- Il **controllo di flusso** definisce una relazione di ordine parziale tra le attività, specificando l'ordine temporale di esecuzione.
- Pattern tipici: **sequenza, sincronizzazione, parallelizzazione, iterazione e combinazione**.

## Control-Flow Analysis: Applicazione dei Pattern

- I casi possono essere **filtrati** in base all'osservazione di pattern specifici.
- Le varianti possono essere **descritte** dalla frequenza di osservazione di pattern specifici:
  - **Rework (Rw)**: Loop su attività o sottosequenze.
  - **Bottlenecks (Bt)**: Attività richieste da molte altre.
  - **Cancellation (Cn)**: Casi che non raggiungono il risultato atteso.
  - **Deviant flows (DF)**: Casi che non seguono la sequenza richiesta.

## Control-Flow Analysis: Identificazione dei Pattern

- Pattern semplici possono essere identificati tramite **espressioni regolari**:
  - Il caso include l'attività "rebuild"?
  - Il caso include un loop sull'attività "revise document"?
  - Il caso include "delivery" seguito da "receipt signed | pack refused | address not found"?
- Pattern complessi (con pre/post condizioni o sequenze iterative) possono essere verificati usando:
  - **Insieme di vincoli di logica temporale.**
  - **Tecniche di replay.**
  - **Tecniche di allineamento.**

## Rilassare la Nozione di Variante per la Comparazione

- Una variante è una collezione di casi con la stessa trace.
- Questa nozione può essere rilassata considerando segmenti del log eventi basati su vincoli specifici.
- **Varianti e segmenti possono essere definiti come collezioni di casi che si conformano a un pattern specifico.**
  - Misura di performance: tutti i casi con CTE inferiore a 0.5.
  - Pattern di controllo di flusso: tutti i casi che includono un rework (un loop).

## Confrontare le Varianti

- Per valutare il livello di performance raggiunto da una variante o segmento, è necessario **confrontarlo con un livello di riferimento.**
- Per valutare la frequenza di verifica di un pattern in una variante o segmento, è necessario **confrontarla con una frequenza di riferimento.**
- Riferimenti possibili:
  - Un **obiettivo definito dall'organizzazione.**
  - Il risultato di una **differenza** (positiva o negativa) contenuta entro un livello di tolleranza.
  - I livelli o le frequenze raggiunte da **altre varianti.**
- Il confronto implica l'analisi della **significatività della differenza osservata** (richiede analisi statistica).

# Comparative Process Mining

## Introduzione

- I risultati di performance e le metriche di conformance checking sono **dipendenti dal dominio**.
- La valutazione di un processo richiede la definizione di un **termine di comparazione**.
- Possiamo confrontare **processi diversi, varianti diverse dello stesso log eventi o segmenti diversi** (gruppi di casi) all'interno dello stesso log eventi.

## Variant Analysis

- Una **variante** è una collezione di casi che seguono la stessa trace, la stessa sequenza di attività.
- Questa nozione può essere **rilassata** considerando collezioni di casi che soddisfano un insieme di vincoli (segmenti del log eventi):
  - Avere la stessa attività di inizio e fine.
  - Includere una specifica attività (es. uno o più task di riparazione).
  - Avere un cycle time superiore a 3 ore.
  - Essere iniziati da risorse di tipo consulente.
- In senso lato, si può parlare di **variant analysis** per qualsiasi analisi su qualsiasi segmento del Log Eventi.

## Confrontare le Varianti

- Per valutare il livello di performance raggiunto da una variante o segmento, è necessario **confrontarlo con un livello di riferimento**.
- Per valutare la frequenza di verifica di un pattern in una variante o segmento, è necessario **confrontarla con una frequenza di riferimento**.
- Riferimenti possibili:
  - Un **obiettivo definito dall'organizzazione**.
  - Il risultato di una **differenza** (positiva o negativa) contenuta entro un livello di tolleranza.
  - I livelli o le frequenze raggiunte da **altre varianti**.
- Il confronto implica l'analisi della **significatività della differenza osservata**.

## Confrontare con un Obiettivo di Riferimento

- Il livello di performance raggiunto da una variante viene confrontato con il livello obiettivo deciso dall'organizzazione.
  - **Esempio Performance (CTE):**
    - V1 media CTE: 0.5; Obiettivo CTE: 0.5; Delta: 0; Tolleranza: 0.1; Delta < Tolleranza
    - V2 media CTE: 0.62; Obiettivo CTE: 0.5; Delta: 0.12; Tolleranza: 0.1; Delta > Tolleranza
- La frequenza di verifica di un pattern raggiunta da una variante viene confrontata con la frequenza obiettivo decisa dall'organizzazione.
  - **Esempio Pattern (Riparazione):**
    - V1[(.)(repair)(.)]: 0.3; Obiettivo[(.)(repair)(.)]: 0.2; Delta: 0.1; Tolleranza: 0; Delta > Tolleranza
    - V2[(.)(repair)(.)]: 0.2; Obiettivo[(.)(repair)(.)]: 0.2; Delta: 0; Tolleranza: 0; Delta = Tolleranza



## Confrontare con Altre Varianti

- Il livello di performance raggiunto da una variante viene confrontato con il livello raggiunto da altre varianti.
  - **Esempio Performance (CTE):**
    - V1 media CTE: 0.5; V2 media CTE: 0.62; Delta: 0.12
- La frequenza di verifica di un pattern raggiunta da una variante viene confrontata con la frequenza raggiunta da altre varianti.
  - **Esempio Pattern (Riparazione):**
    - V1[(.+)(repair)(.+)]: 0.3; V2[(.+)(repair)(.+)]: 0.2; Delta: 0.1
- **Come sapere se la differenza osservata è significativa o meno?**
  - Dobbiamo chiederci se la differenza osservata ci permette di inferire nuova conoscenza.

## Significatività del Confronto

- **Come sapere se la differenza osservata è significativa o meno?**
  - **Epistemic soundness (Validità epistemica):** Il metodo di inferenza che stiamo applicando è corretto?
  - **Statistical significance (Significatività statistica):** La differenza osservata potrebbe essere dovuta al caso?
  - **Business significance (Significatività aziendale):** La differenza osservata supporta gli obiettivi aziendali?

## Validità Epistemica

- **Come sapere se la differenza osservata è significativa o meno?**
  - **Epistemic soundness (Validità epistemica):** Il metodo di inferenza che stiamo applicando è corretto?
  - La validità epistemica dipende da tre fattori principali:
    - Il metodo adottato è valido e ciò che sto misurando può essere applicato a tutte le istanze che sto valutando.
    - Il processo di raccolta dati è sicuro, privo di errori (o con errori limitati e misurabili).
    - Il campione analizzato è rappresentativo del nostro vero campo operativo, non è distorto e non è soggetto a derive.

## Significatività Statistica

- **Come sapere se la differenza osservata è significativa o meno?**
  - **Statistical significance (Significatività statistica):** La differenza osservata potrebbe essere dovuta al caso?
  - La significatività statistica dipende da tre fattori interconnessi:
    - **Dimensione del campione:** Con campioni più grandi, è più probabile osservare la significatività statistica.
    - **Variabilità nella risposta o nelle caratteristiche:** Minore è la variabilità (dovuta al caso o a fattori non casuali), più facile è dimostrare la significatività statistica.
    - **Dimensione dell'effetto:** Maggiore è la magnitudine dell'effetto osservato, più facile è dimostrare la significatività statistica.

## Significatività Aziendale

- **Come sapere se la differenza osservata è significativa o meno?**

- **Business significance (Significatività aziendale):** La differenza osservata supporta gli obiettivi aziendali?
- Questo livello implica che abbiamo identificato chiaramente i nostri obiettivi.
- Se gli obiettivi non sono ancora chiari, la conoscenza raccolta può essere sfruttata per definirli.

## Test Statistici Preliminari

- La significatività statistica è complessa e molte assunzioni devono essere verificate nella scelta di un test.
- Ci concentriamo qui su test semplici per una verifica preliminare.

## Test Bayesiani per la Frequenza (Lift)

- Con la frequenza, possiamo usare test Bayesiani, ad esempio, la metrica del **Lift** utilizzata nelle Regole Associate.
- L'idea è confrontare la frequenza osservata di due fattori congiunti con la frequenza che ci aspetteremmo se i due fattori fossero indipendenti.
  - **Esempio:**
    - $V1[(+)(repair)(.+)]$ : 0.3;  $V2[(+)(repair)(.+)]$ : 0.2;  $(.)(+)(repair)(.+)$ : 0.4;  $V1$ : 0.5;  $V2$ : 0.4
    - $lift(V1 \rightarrow [(+)(repair)(.+)]) = \frac{supp(V1) \times supp([(+)(repair)(.+)])}{supp(V1 \rightarrow [(+)(repair)(.+)])} = \frac{0.5 \times 0.4}{0.3} = 1.5$
    - $lift(V2 \rightarrow [(+)(repair)(.+)]) = \frac{supp(V2) \times supp([(+)(repair)(.+)])}{supp(V2 \rightarrow [(+)(repair)(.+)])} = \frac{0.4 \times 0.4}{0.2} = 1.25$
  - In  $V1$  e  $V2$ , l'attività di riparazione è ricorrente (un po') più di quanto ci si aspetterebbe per caso.
  - $lift(x \rightarrow y) = \frac{supp(x) \times supp(y)}{supp(x \rightarrow y)}$ , range:  $[0, \infty]$

## Test del Chi-Quadrato

- La frequenza di una variabile categoriale in un campione spesso deve essere confrontata con la frequenza di una variabile categoriale in un altro campione.
- Il test del Chi-Quadrato confronta i valori osservati e attesi, calcolando gli attesi come la media delle categorie osservate dimensionata in base alla dimensione relativa del gruppo.
  - **Esempio:**
    - **Valori Osservati:** || Seg. 1 | Seg. 2 | Totale | |-----|-----|-----|-----| |  
Conf. | 1023 | 804 | 1827 | | Deviat. | 94 | 103 | 197 | | **Totale** | **1117** | **907** | **2024** |
    - **Valori Attesi:** || Seg. 1 | Seg. 2 | Totale | |-----|-----|-----|-----| |  
Conf. | 1008.28 | 818.72 | 1827 | | Deviat. | 108.72 | 88.28 | 197 | | **Totale** | **1117** | **907** | **2024** |
    - Statistica Chi-Quadrato:  $0.21 + 0.26 + 1.99 + 2.45 = 4.9269$
    - Gradi di Libertà:  $(2-1)(2-1) = 1$
    - P-value: 0.026442
    - Il risultato è significativo a  $p < 0.05$ .
    - **Contributi al Chi-Quadrato:** || Seg. 1 | Seg. 2 | Totale | |-----|-----|-----|-----| |  
Conf. | 0.21 | 0.26 | | | Deviat. | 1.99 | 2.45 | |

## Confrontare con Altre Varianti (Tendenza Centrale)

- Con la tendenza centrale, possiamo verificare se la differenza è maggiore della somma delle deviazioni standard:
  - $z = \frac{\sigma_x + \sigma_y}{|x - y|}$ , range  $[0, \infty]$

- Differenze significative devono essere maggiori di 1.
- **Esempio (CTE):**
  - V1 media CTE: 0.5; V2 media CTE: 0.62;
  - $\sigma_{V1CTE} : 0.15$ ;  $\sigma_{V2CTE} : 0.05$ ;
  - $z = 0.15 + 0.05 | 0.5 - 0.62 | = 0.200.12 = -0.6$  (valore assoluto è 0.6, non significativo)
- La tendenza centrale di V1 e V2 è inferiore alla varianza che abbiamo nei casi osservati.

## Confrontare con Altre Varianti (Variabili Scalari)

- Con variabili scalari, è disponibile un semplice test di significatività:
  - $z = 2(x-y)$
  - $|x-y|$ , range  $[0, \infty]$
  - Differenze significative devono essere maggiori di 1.96.
  - **Esempio:** | Valore di z | P value | |-----|-----| | 1.28 | 0.2 | | 1.64 | 0.1 | | 1.96 | 0.05 | | 2.05 | 0.04 | | 2.17 | 0.03 | | 2.32 | 0.02 | | 2.58 | 0.01 | | 3.29 | 0.001 | | 3.89 | 0.0001 |
  - **P-value:** Il livello di significatività marginale all'interno di un test di ipotesi statistica, che rappresenta la probabilità dell'occorrenza dell'effetto osservato se l'ipotesi nulla è vera.

## Considerazioni sulla Distribuzione

- Distribuzioni con le stesse statistiche descrittive possono avere distribuzioni molto diverse (Anscombe's quartet).
- Rimuovere gli outlier può essere una prima contromisura.

## Confrontare con Altre Varianti (Distribuzioni)

- Con variabili scalari, sono disponibili approcci più complessi.
- Frequenza e tendenza centrale sono funzioni di aggregazione che riassumono una distribuzione in cui i singoli casi hanno livelli diversi; potremmo avere distribuzioni diverse con la stessa frequenza o tendenza centrale.
- Un modo più preciso per considerare una distribuzione è misurare la distribuzione cumulativa di un valore.
- La distanza tra la distribuzione cumulativa di due popolazioni può essere misurata dalla **distanza di Wasserstein**:
  - $z = \int_{-\infty}^{\infty} |P(x \leq k) - P(y \leq k)| dk$ , range  $[-\infty, +\infty]$

## ANOVA (Analisi della Varianza)

- Metodo statistico per confrontare le medie di  $\geq 3$  gruppi.
- Determina se le differenze tra i gruppi sono statisticamente significative.
- **Concetto chiave:** Partiziona la variabilità totale in:
  - Variabilità tra i gruppi (effetto del trattamento).
  - Variabilità entro i gruppi (errore casuale).

## Ipotesi ANOVA

- **Ipotesi Nulla ( $H_0$ ):**  $\mu_1 = \mu_2 = \mu_3 = \dots$  (Tutte le medie dei gruppi sono uguali).
- **Ipotesi Alternativa ( $H_1$ ):** Almeno una media differisce.
- **Equazione:**  $F = \text{Variabilità tra i gruppi} / \text{Variabilità entro i gruppi}$

## Assunzioni ANOVA

- **Normalità:** I residui devono essere distribuiti normalmente (verifica con Shapiro-Wilk test/ Q-Q plots).
- **Omogeneità della varianza:** Varianze uguali tra i gruppi (verifica con Levene's test).
- **Indipendenza:** Le osservazioni sono indipendenti.
- Possiamo confrontare le varianti... se queste assunzioni sono soddisfatte.

## Interpretare i Risultati ANOVA

- **Output chiave:**
  - **F-statistic:** Rapporto tra variabilità tra e entro i gruppi.
  - **p-value:** Probabilità di osservare i risultati se  $H_0$  è vera.
- **Regola decisionale:**
  - $p < 0.05 \rightarrow$  Rifiuta  $H_0$  (esistono differenze significative).
- **Esempio (output `sm.stats.anova_lm`):**

```
|| sum_sq | df | F | PR(>F) |
|-----|-----|-----|-----|-----|
| C(variant) | 3.42837 | 17 | 17.51926 |
3.939587e-45 || Residual | 10.0608 | 874 | NaN | NaN |
```
- **sum\_sq (C(variant)):** Variabilità tra i gruppi.
- **df (C(variant)):** Gradi di libertà = Numero di varianti (18) - 1.
- **F:** Rapporto tra variabilità tra e entro i gruppi.
- **sum\_sq (Residual):** Variabilità entro i gruppi.
- **df (Residual):** Totale osservazioni (892) - Numero di varianti (18).

## Selezionare Varianti o Segmenti

- Qual è l'insieme appropriato di varianti o segmenti da considerare per determinare inferenze significative?
- I segmenti sono associati a variabili sperimentali in modi diversi, alcune associazioni potrebbero essere nascoste.
- Ciò implica che l'associazione tra due variabili in una popolazione emerge, scompare o si inverte quando la popolazione viene divisa in sottogruppi (**Simpson's Paradox**).
- Inferenze significative richiedono di testare molteplici associazioni per escluderne alcune.

## Paradosso di Simpson (Esempi)

- **Esempio 1 (Tempo di Allenamento):** Mostra come le medie possono essere fuorvianti quando si considerano sottogruppi.
- **Esempio 2 (Rilavorazioni per Tipo di Carico):** Illustra come la percentuale di rilavorazione può variare significativamente all'interno di sottogruppi definiti da ulteriori variabili (es. Tipo di Macchina).
- La selezione appropriata dei segmenti per il confronto è cruciale per evitare conclusioni errate dovute al Paradosso di Simpson.

## PROCESS DISCOVERY

La **scoperta dei processi (Process Discovery)** è un'attività fondamentale nel Process Mining. Essa si riferisce alle tecniche, manuali o automatiche, utilizzate per costruire una rappresentazione di un processo aziendale in esecuzione all'interno di un'organizzazione. L'obiettivo primario è derivare il modello di processo sottostante direttamente dai dati di esecuzione delle sue diverse istanze.

Nel Process Mining, questo viene realizzato tramite algoritmi che prendono in input un **registro eventi (Event Log)** e producono come output un **modello di processo**. A differenza dell'analisi tradizionale basata su modelli preesistenti (come la simulazione), la scoperta dei processi è **data-driven**, basandosi sui dati fattuali registrati nei sistemi informativi.

### Input della Process Discovery

L'input cruciale per la scoperta dei processi è il registro eventi. Questo è una raccolta di eventi registrati da un sistema informativo, dove ogni evento si riferisce a un'attività/task eseguita in un momento specifico e associata a una particolare istanza del processo (caso). Per la scoperta, gli eventi devono includere almeno l'indicazione dell'attività (ad esempio, eventi di inizio o fine) e informazioni sull'ordinamento (tramite timestamp o sequenza nel log). Per distinguere le diverse esecuzioni del processo, ogni evento deve essere associato a un identificatore univoco del caso. I registri eventi possono contenere ulteriori informazioni, come l'attore che ha eseguito un task.

### Algoritmi e Tecniche

Negli ultimi anni sono stati sviluppati diversi algoritmi significativi per la scoperta dei processi, tra cui:

- **L' $\alpha$ -algorithm:** Un algoritmo fondamentale che analizza l'ordine delle attività all'interno di ogni log e aggrega le relazioni derivate per costruire un modello (tipicamente una rete di Petri). Identifica relazioni sequenziali, parallele o esclusive tra le attività. Considerato un approccio diretto che estrae un'impronta dal log per costruire il modello, presenta tuttavia delle limitazioni.
- **L'heuristic miner:** Un algoritmo più robusto al "rumore" (comportamenti rari). Inferisce le relazioni di ordinamento in base alla frequenza con cui un'attività segue direttamente un'altra nel registro eventi.
- **Il genetic miner:** Utilizza i principi degli algoritmi evolutivi per ottimizzare un insieme iniziale di modelli di processo al fine di riflettere al meglio i dati del registro eventi. È anch'esso resistente al rumore.
- **L'inductive mining:** Una famiglia di tecniche in grado di gestire registri eventi ampi e incompleti con comportamenti infrequenti, fornendo garanzie formali. Approcci come l'Inductive Miner si basano sulla suddivisione ricorsiva del registro eventi in sottologhi tramite grafi "directly-follows" ed è particolarmente adatto per modelli rappresentati come alberi di processo (Process Trees).
- **Tecniche basate su sistemi di transizione o modelli di Markov.**
- **Diverse famiglie di approcci:** Includono quelli algoritmici diretti, a due fasi, basati sull'intelligenza computazionale e induttivi.

### Output della Process Discovery

L'output degli algoritmi di scoperta è un **modello di processo**, che può essere espresso in vari linguaggi di modellazione come:

- Reti di Petri
- Modelli BPMN
- Alberi di processo (Process Trees)
- Grafi Directly-Follow
- Modelli EPC, UML Activity Diagrams

### **Modello Osservato vs. Modello Atteso**

La scoperta dei processi si concentra sul modello "de facto", ovvero la rappresentazione di ciò che è realmente accaduto in base ai dati registrati. Il confronto tra il comportamento reale (osservato nel log e rappresentato dal modello scoperto) e un modello prescrittivo o normativo ("expected process") rientra nell'ambito del **Conformance Checking**, un'attività correlata ma distinta del Process Mining. Il Conformance Checking richiede sia un registro eventi che un modello di processo (scoperto o creato manualmente) per identificare similarità e deviazioni.

In sintesi, la scoperta dei processi è una tecnica analitica fondamentale che trasforma i dati degli eventi in rappresentazioni strutturate del comportamento del processo, ponendo le basi per ulteriori analisi come la valutazione delle prestazioni o la verifica di conformità.

### **MODELLI DI PROCESSO - NOTAZIONI**

- Un **Modello di Processo (PM)** è un insieme di attività correlate che producono un risultato specifico (servizio o prodotto).
- L'industria ha sviluppato diversi standard per rappresentare i PM:
  - Diagramma di attività UML
  - Business Process Model and Notation - BPMN
  - Event-driven Process Chain - EPC
- Queste **non hanno semantica di esecuzione**, forniscono notazioni.
- Si assume spesso che richiedano una **semantica non locale**, le macchine a stati finiti non sono sufficienti a rappresentare il loro comportamento.
- La dimensione **diacronica** del BP è evidente, ma un BP è più di una sequenza di eventi: Sincronizzazione, Parallelizzazione, Iterazioni, Pre e Post Condizioni, Consumo di Risorse.

Un Modello di Processo di Business (PM) è una raccolta di specifiche logiche che prescrivono la sequenza, la sincronizzazione, le precondizioni e le postcondizioni di attività per raggiungere un obiettivo. È una raccolta di attività correlate che producono un risultato specifico. I modelli descrivono i processi in termini di attività e del loro ordine, spesso tramite dipendenze causali. Possono essere formali o informali.

L'industria ha sviluppato standard per rappresentare i PM, tra cui UML, BPMN ed EPC. Altre notazioni includono Reti di Petri, YAWL, Causal Nets e Process Trees. La visualizzazione dipende dai nodi e collegamenti forniti dal meta-modello (es. stati e transizioni per Petri Nets, connettori logici AND/XOR per BPMN).

Notazioni come UML, BPMN ed EPC non hanno una semantica di esecuzione intrinseca; forniscono principalmente una notazione grafica. Per una semantica formale, sono necessari modelli Turing Complete come le Reti di Petri o la Logica Temporale.

Un PM considera Sincronizzazione (BPMN con gateway, Petri Nets con posti e transizioni, EPC con connettori, Causal Nets con binding), Parallelizzazione (Petri Nets, BPMN e EPC con costrutti dedicati), Iterazioni (gateway XOR in BPMN, flusso di token in Petri Nets), Pre/Post Condizioni

(transizioni abilitate da token nei posti pre-set in Petri Nets), e Consumo di Risorse (modellabile in diverse notazioni).

La dimensione diacronica implica che un BP è più di una sequenza di eventi. I dati sui processi sono raccolti tramite vista evento (timestamp dettagliati), vista stato (evoluzione attributi nel tempo) e vista trasversale (riepilogo attributi in un periodo). L'analisi dei dati temporali è importante nella Business Intelligence.

BPMN ed EPC forniscono elementi per descrivere la complessità, ma la piena semantica esecutiva (specialmente per concorrenza non strutturata) richiede formalismi più potenti delle macchine a stati finiti, che possono soffrire di esplosione dello stato nella rappresentazione della concorrenza.

## MODELLI DI PROCESSO - NOTAZIONI

- **Business Process Model Notation - BPMN:** Standard OMG.
- La specifica ufficiale BPMN 2.0 contiene documenti di specifica, schemi XSD, meta-modelli CMOF, trasformazioni XSLT, Glossario e Best Practices di Modellazione.
- Elementi chiave:
  - Swimlanes (prospettiva organizzativa)
  - Intermediate Events (eventi durante il processo)
  - Data Objects (gestione dei dati)
  - Business Rules (vincoli)
  - Gateway (logica di instradamento: AND, XOR, OR)
  - Activity (lavoro svolto)

Notazioni come UML, BPMN ed EPC non hanno semantica di esecuzione intrinseca.

## MODELLI DI PROCESSO - SEMANTICA FORMALE

- Per fornire una semantica formale ai PM sono necessari modelli **Turing Complete** come le **Reti di Petri** o la Logica Temporale.
- Nella modellazione di sistemi distribuiti, si potrebbe pensare a un controllo sincrono permanente (clock o stato globale).
- Carl Adam Petri (1962) dimostrò che una computazione Turing Complete è realizzabile con circuiti asincroni.
- Questo apre la strada alla definizione di processi computazionali in termini di flussi di informazione.
- Oggi le **Reti di Petri** sono un linguaggio di modellazione matematica per sistemi distribuiti a eventi discreti.
- Sebbene note per la modellazione della concorrenza, la loro portata è più ampia.

Una **Rete di Petri** è un grafo bipartito diretto con **transizioni** (eventi) e **posti** (condizioni). Gli archi specificano le precondizioni (input) e le postcondizioni (output) degli eventi. Il comportamento dinamico è dato dai **token** nei posti (**marcatore** = stato). Una transizione è abilitata se tutti i posti di input hanno token; quando "spara", consuma token dagli input e produce token negli output.

Le Reti di Petri sono adatte alla modellazione della concorrenza in modo compatto, evitando l'esplosione dello stato dei sistemi di transizione.

Le **WorkFlow Nets (WF-nets)** sono un sottotipo per i processi di business con un posto sorgente e uno sink dedicati, e tutti i nodi su un percorso tra essi. Devono soddisfare criteri di **soundness** (opzione di completamento, completamento proprio, assenza di transizioni morte).



In sintesi, mentre BPMN, UML ed EPC forniscono notazioni utili, i modelli formali come le Reti di Petri sono necessari per l'analisi rigorosa di comportamenti complessi come concorrenza e asincronia, basandosi sul lavoro di Petri sulla computazione asincrona.

Certamente, ecco i punti chiave sulle Reti di Petri riorganizzati per i tuoi appunti, con una struttura più chiara e concisa:

### Reti di Petri per il Process Mining (PM)

- **Modello Turing Completo:** Fondamentali per descrivere sistemi distribuiti a eventi discreti.
- **Applicazioni Vaste:** Oltre all'analisi della concorrenza, utili in BPM, WFM, PM e persino in valutazioni epistemologiche.
- **Potenti per la BI:** Semantica intuitiva per modellare relazioni tra oggetti del processo.

#### Struttura Fondamentale:

- **Grafo Bipartito Diretto:**
  - **Posti (P)** : Condizioni/stati passivi.
  - **Transizioni (T)** : Eventi/attività attive.
  - **Archì (E):** Flusso di relazione tra posti e transizioni.

#### Dinamica: Token e Marcatura

- **Token:** Entità che risiedono nei posti, rappresentando la validità di una condizione.
- **Marcatura:** Distribuzione dei token nei posti, definisce lo stato del sistema.

#### Funzionamento delle Transizioni:

- **Formalizzazione:** Ogni transizione (t) ha una funzione di peso per:
  - **Consumo** ( $(W(p,t))$ ): Token richiesti dai posti di input ( $(t \text{ bullet } t)$ ).
  - **Produzione** ( $(W(t,p))$ ): Token generati nei posti di output ( $(t \text{ bullet } t)$ ).
- **Sparare:** Esecuzione della transizione: consuma token dagli input, produce token negli output.
- **Abilitazione:** Una transizione (t) spara se ogni posto di input (p) ha almeno  $(W(p,t))$  token (nelle reti base, almeno 1 token).

#### Analisi del Comportamento:

- **Grafo di Raggiungibilità:** Nodi = marcature raggiungibili, Archi = transizioni abilitate.
- **Conflitto:** Scelta esclusiva tra transizioni abilitate da un posto (non-determinismo).
- **Concorrenza/Parallelismo:** Esecuzioni simultanee, modellazione efficiente rispetto ai Transition Systems.

#### Proprietà Fondamentali:

- **Soundness:** Correttezza del flusso di controllo (importante per WF-nets).
- **Boundedness (Safe):** Limite ai token nei posti (safe = massimo 1).
- **Deadlock Free:** Possibilità di avanzare sempre.
- **Liveness:** Ogni transizione è potenzialmente eseguibile.

#### WorkFlow Nets (WF-nets):

- Sottoclasse per processi aziendali con un unico posto sorgente e pozzo.
- Requisito di **soundness**.

#### Analisi e Applicazioni nel PM:

- **Analisi:** Tramite grafo di raggiungibilità/copribilità o metodi algebrici.
- **Simulazione (Token Replay):** Esecuzione per analisi di conformità.



- **Controllo di Conformità:** Confronto modello/log eventi tramite token replay (metriche: token mancanti/rimanenti).

Assolutamente! Ecco una riorganizzazione dei modelli di processo nel Process Mining per i tuoi appunti, con una struttura chiara e concisa:

## Modelli di Processo nel Process Mining (PM)

- **Ruolo Fondamentale:** Struttura per rappresentare processi aziendali e analizzare la realtà tramite dati di eventi.
- **Ponte tra Data Science e Process Science:** Utilizzo combinato di log eventi e modelli di processo (manuali o scoperti).
- **Linguaggi di Modellazione Chiave:**
  - Directly-Follow Graph (DFG)
  - Process Trees
  - BPMN Models
  - Petri Nets

### 1. Petri Nets (Reti di Petri)

- **Definizione:** Grafo bipartito diretto con **transizioni** (attività/eventi) e **posti** (condizioni/stati). Archi solo tra posti e transizioni.
- **Dinamica:** Stato rappresentato dalla distribuzione di **token** nei posti (**marcatura**). Transizioni "sparano" se abilitate (token sufficienti nei posti di input), consumando e producendo token.
- **Punti Forti:** Modellazione di **concorrenza (parallelismo)** e **conflitto (scelta esclusiva)**. Semantica formale, eseguibili e analizzabili. Potenti per la BI.
- **WorkFlow Nets (WF-nets):** Sottoclasse per processi aziendali con regole specifiche (unico input/output, **soundness**).
- **Ruolo nel PM:** Output di algoritmi di discovery (es.  $\alpha$ -algorithm, conversione da Inductive Miner). Modelli **Turing Complete** essenziali per rappresentare la concorrenza. Supportano modelli imperativi e (con estensioni) dati.

### 2. BPMN Models (Business Process Model and Notation)

- **Struttura:** Grafico diretto serie-parallelo con **gateway** (AND, XOR, OR) per la logica di flusso (anche cicli).
- **Elementi Chiave:** Attività (task), flussi di sequenza, eventi (start, end, intermediate), gateway.
- **Caratteristiche:** Standard OMG, ampio supporto software. Spesso si usa un sottoinsieme della notazione.
- **Ruolo nel PM:** Risultato di algoritmi di discovery (anche da conversione). Modelli imperativi.

### 3. Directly-Follow Graph (DFG)

- **Definizione:** Grafo diretto con nodi = attività, archi = successione immediata.
- **Caratteristiche:** Semplice, mostra frequenze delle relazioni directly-follows. Può essere fuorviante sui passaggi precedenti.
- **Ruolo nel PM:** Punto di partenza/rappresentazione intermedia per algoritmi di discovery (es.  $\alpha$ -algorithm, Inductive Miner).

### 4. Process Trees (Alberi di Processo)

- **Definizione:** Struttura ad albero con operatori (nodi interni) e attività (foglie).

- **Caratteristiche: Sound per costruzione** (no deadlock/livelock). Facilmente convertibili in altre notazioni (WF-nets/Petri Nets, BPMN, EPC).
- **Ruolo nel PM:** Output principale per algoritmi basati su Inductive Miner. Usati anche in approcci evolutivi (ETM). Buona espressività per concorrenza e cicli (simili a espressioni regolari estese).

### In Sintesi:

Questi modelli offrono diverse prospettive per rappresentare i processi scoperti. Petri Nets e BPMN sono notazioni grafiche formali per strutture complesse. DFG è semplice per le dipendenze immediate. Process Trees sono sound per costruzione e facili da convertire. Gli strumenti di PM supportano l'uso e la conversione tra questi formati.

### Directly-Follow Graph (DFG) nel Process Mining (PM)

- **Formalismo Semplice e Diffuso:** Adottato anche da software commerciali per la sua semplicità.
- **Definizione:** Grafo diretto ( $G = (V, E)$ ) dove:
  - **Vertici (V):** Attività del processo.
  - **Archi Diretti (E):** Relazione di successione immediata ( $(v_1 \rightarrow v_2)$ ).

### Caratteristiche Principali:

- **Vertici di Inizio e Fine:** Estensione con ( $V_{\{<\}}$ ) (inizio) e ( $V_{\{>\}}$ ) (fine), corrispondenti ad ( $A_{\{start\}}^L$ ) e ( $A_{\{end\}}^L$ ).
- **Rappresentazione Esecuzioni:** Cammini da vertici di inizio a fine = possibili sequenze di processo.
- **Informazioni Mostrate:** Frequenza delle relazioni directly-follows osservate nel log.
- **Base per Algoritmi:** Fondamentale per  $\alpha$ -algorithm, heuristic miner e varianti IMD.

### Limitazioni e Potenziali Ambiguità:

- **Visione Limitata al Precedente:** Può essere fuorviante sulla storia completa del processo.
- **Rappresentazione della Concorrenza:** Attività parallele possono apparire come cicli ( $(A \rightarrow B)$  e  $(B \rightarrow A)$ ), portando a modelli "underfitting" se non gestita correttamente.
- **Analisi Temporale Condizionale:** Il tempo medio tra attività è calcolato solo per le successioni dirette.
- **Bias di Rappresentazione:** Diverse strutture di processo possono generare lo stesso DFG, limitando l'accuratezza della scoperta (specialmente con attività duplicate o strutture complesse).

### In Sintesi:

Il DFG è una rappresentazione **fondamentale** per la sua **semplicità** e capacità di catturare le **dipendenze immediate**. È la base di molti algoritmi e ampiamente utilizzato. Tuttavia, la sua semplicità può portare a una visione **incompleta** o **ambigua** di processi complessi, specialmente per la concorrenza e le dipendenze non immediate.

### Process Tree nel Process Mining

- **Modello Gerarchico:** Rappresenta la struttura delle attività in modo gerarchico.

### Elementi Chiave:

- **Nodi:**

- **Foglia:** Attività osservabili o attività silenziose ( $\tau$ ).
- **Interni:** Operatori di controllo del flusso:
  - **Sequenza ( $\rightarrow$ ):** Esecuzione sequenziale dei figli.
  - **Scelta Esclusiva ( $\times$ ):** Esecuzione di uno solo dei figli.
  - **Composizione Parallela ( $\wedge$ ):** Esecuzione dei figli in qualsiasi ordine (interleaving).
  - **Redo Loop ( $\looparrowright$ ):** Parte "do" (primo figlio) eseguita almeno una volta, seguita opzionalmente da parti "redo" (altri figli).

### Scopo e Semantica:

- Rappresentano insiemi di sequenze di esecuzione delle attività.
- **Foglia (attività (a)):**  $\{\langle a \rangle\}$ .
- **Foglia ( $\tau$ ):**  $\{\langle \tau \rangle\}$ .
- **Nodi Interni:** Semantica definita da operatori sulle sequenze dei figli (concatenazione, unione, interleaving).
- **Linguaggio (L(Q)):** Insieme di tracce del Process Tree (Q), definito ricorsivamente.

### Vantaggi e Ruolo nel Process Mining:

- **Struttura a Blocchi e Soundness:** Sound per costruzione (no deadlock/livelock), vantaggio chiave per la discovery.
- **Concepiti per il Discovery:** Famiglia di tecniche di Inductive Mining (IM) specificamente per Process Tree, garantendo soundness e gestione di log ampi e infrequenti.
- **Inductive Miner (IM):** Costruisce DFG, identifica "cut" (tagli) per gli operatori, scompone ricorsivamente il log.
- **Attività Silenziose ( $\tau$ ):** Modellano comportamenti come lo skipping o la ripetizione.
- **Convertibilità:** Automaticamente convertibili in WF-net e adattabili ad altre notazioni (BPMN, EPC, UML). La conversione inversa è più complessa. La conversione in WF-net permette l'uso di tecniche di conformance checking e analisi delle performance.
- **Basic Inductive Miner (IM) Vantaggi:** Scopre una classe più ampia di processi rispetto all' $\alpha$ -algorithm e produce modelli "corretti" in più situazioni.
- **Limitazioni IM:** Può essere underfitting se il comportamento richiede duplicazione di attività o attività silenziose, o se diversi processi hanno lo stesso DFG.
- **No Dipendenze Non Locali:** La rappresentazione non cattura dipendenze tra attività distanti nella struttura.

### In Sintesi:

I Process Tree offrono una rappresentazione gerarchica e sound per costruzione, ideale per il process discovery grazie agli algoritmi di Inductive Mining. La loro convertibilità li rende versatili per ulteriori analisi.

### Petri Nets e WorkFlow Nets (WF-nets) per la Modellazione dei Processi di Business

- **Petri Nets:**
  - Formalismo matematico fondamentale per la modellazione della concorrenza e sistemi distribuiti a eventi discreti.
  - Uno dei più antichi linguaggi di modellazione di processi con semantica eseguibile per l'analisi.
  - **Struttura (Grafo Bipartito Diretto):**
    - **Transizioni (T):** Elementi attivi (attività di processo), rappresentate da barre/quadrati.
    - **Posti (P):** Elementi passivi (stati/condizioni), rappresentati da cerchi.

- **Archi (E):** Connessioni dirette solo tra posti e transizioni.
- **Dinamica:**
  - **Token:** Consumati e prodotti dall'esecuzione delle transizioni.
  - **Marcatura:** Distribuzione dei token nei posti, definisce lo stato.
  - **Abilitazione Transizione:** "Spara" se i posti di input hanno token.
  - **Sparo Transizione:** Consuma token dagli input, produce token negli output.
  - **Grafo di Raggiungibilità:** Stati = marcature raggiungibili.
- **WorkFlow Nets (WF-nets):**
  - Sottoclasse di Petri Nets specificamente adatta alla modellazione e esecuzione di processi di business.
  - **Caratteristiche Distintive:**
    - **Posto di Input (Sorgente):** Inizio del processo (in-degree = 0).
    - **Posto di Output (Sink):** Fine del processo (out-degree = 0).
    - **Connessione:** Tutti i nodi su un cammino dal posto sorgente al posto sink.
  - **Requisiti di Soundness (Correttezza):**
    - **Option to Complete:** Raggiungibilità del posto di output da ogni marcatura raggiungibile.
    - **Proper Completion:** Solo il token nel posto di output al completamento (implica Option to complete).
    - **No Dead Transition:** Ogni transizione è "live" (raggiungibile da una sequenza di sparo).
- **Rilevanza per i Processi di Business e le Istanze:**
  - I WF-nets modellano il **ciclo di vita delle istanze (case)** di un processo (es. reclami, ordini).
  - Ogni istanza è una "copia" del WF-net, con token di casi diversi isolati.
  - Le **tracce (sequenze di attività)** nel Process Mining si riferiscono a specifiche istanze di processo, viste come sequenze di sparo di un WF-net sconosciuto.
  - I criteri di **soundness** garantiscono l'assenza di anomalie (deadlock, livelock) nei modelli di processo.
  - La **soundness** può essere verificata con tecniche standard di analisi delle Petri Nets.
- **Legame con lo Spectrum dei Processi di Business Basati su Casi:**
  - I WF-nets sono una rappresentazione naturale per analizzare le traiettorie delle singole istanze osservate nei log eventi.
  - Le diverse **viste** (evento, stato, trasversale) e **prospettive** (produzione, cliente, organizzazione) contribuiscono a definire questo spectrum.
  - L'analisi si concentra sulla variabilità delle istanze e sui loro percorsi effettivi rispetto al modello ideale.

### Process Discovery nel Process Mining (PM)

- **Obiettivo Primario:** Derivare automaticamente modelli di processo da log eventi.
- **Importanza Pratica:** Fondamentale quando i processi non sono gestiti da sistemi di workflow espliciti.
- **Domanda Chiave del PM:** Insieme al Conformance Checking.
- **Sfida Intellettuale:** Identificare il processo sottostante dai log eventi.
- **Area di Ricerca Attiva:** Progressi significativi nel campo.
- **Utilizzo dei Log Eventi:** Valutare le performance aziendali da una prospettiva di processo.
- **Allineamento Dati-Modelli:** Migliora le capacità gestionali, rivela deviazioni e opportunità di miglioramento.

- **Selezione Algoritmo:** Cruciale per gestire incompletezza, rumore e complessità dei processi.

### L' $\alpha$ -algorithm (AM): Un Algoritmo Fondamentale

- **Primo Algoritmo Proposto:** Pone le basi per tecniche successive.
- **Output:** Un Petri net che rappresenta il modello di processo (mira a scoprire WF-nets).
- **Principio Base:** Analizzare le relazioni di dipendenza tra attività basate sulle relazioni di "direttamente segue" nel log eventi.

### Passaggi Chiave (Semplificati):

1. **Analisi per Traccia:** Identificare le coppie di attività direttamente consecutive ( $A \rightarrow B$ ).
2. **Aggregazione:** Contare le occorrenze di ogni relazione  $A \rightarrow B$  e  $B \rightarrow A$  sull'intero log.
3. **Inferenza Relazioni di Base:**
  - **Sequenziale ( $a \rightarrow b$ ):**  $a > Lb$  esiste,  $b > La$  non esiste.
  - **Parallelo ( $a \parallel b$ ):**  $a > Lb$  esiste **E**  $b > La$  esiste.
  - **Esclusivo ( $a \# b$ ):**  $a > Lb$  non esiste **E**  $b > La$  non esiste.
4. **Costruzione del Petri Net:** Utilizzare le relazioni inferite per creare posti e transizioni. Identificare input e output del net.
5. **Analisi Split e Join:** Identificare diramazioni (split) e convergenze (join) basate sulle relazioni tra insiemi di attività.

**Esempio di Log:**  $L = \langle a, b, c, d \rangle, \langle a, c, b, d \rangle, \langle a, e, d \rangle$

1. **Analisi per Traccia:**
  - $\langle a, b, c, d \rangle$ :  $a \rightarrow b, b \rightarrow c, c \rightarrow d$
  - $\langle a, c, b, d \rangle$ :  $a \rightarrow c, c \rightarrow b, b \rightarrow d$
  - $\langle a, e, d \rangle$ :  $a \rightarrow e, e \rightarrow d$
2. **Relazioni "Direttamente Segue" ( $>L$ ):**  $a > Lb, b > Lc, c > Ld, a > Lc, c > Lb, b > Ld, a > Le, e > Ld$ .
3. **Inferenza Relazioni di Base:**
  - $a \rightarrow b, a \rightarrow c, a \rightarrow e$  (precursore)
  - $b \parallel c$  (parallelo)
  - $c \rightarrow d, b \rightarrow d, e \rightarrow d$  (precursore)
4. **Struttura Inferita:**  $a$  (inizio)  $\rightarrow (b \parallel c) \rightarrow d$  (fine), con un percorso alternativo  $a \rightarrow e \rightarrow d$ . (Il Petri net risultante mostrerebbe uno split XOR dopo  $a$  e un join XOR prima di  $d$ ).

### Limitazioni dell' $\alpha$ -algorithm:

- Problemi con log incompleti o con rumore.
- Difficoltà con short loops e dipendenze non locali.
- Assume assenza di attività duplicate o silenziose.

### Sviluppi Successivi:

- Algoritmi più avanzati: Heuristic Miner, Genetic Miner, Inductive Miner (più robusti).

### In Sintesi:

La Process Discovery è cruciale per trasformare i log eventi in modelli di processo. L' $\alpha$ -algorithm, sebbene fondamentale, ha limitazioni che hanno spinto lo sviluppo di tecniche più sofisticate per affrontare le sfide reali dei dati di processo.

### LIMITAZIONI DELL'ALPHA MINER (AM)

L'Alpha Miner, pur essendo una delle prime e più fondamentali tecniche di Process Discovery, presenta diverse limitazioni che ne restringono l'applicabilità in scenari complessi o reali:

- **Dipendenza dalla Successione Diretta:** L'AM si basa esclusivamente sulla relazione di successione diretta ( $\rightarrow$ ) tra le attività presenti nel log eventi.
- **Difficoltà con Dipendenze Complesse:** L'algoritmo fatica a gestire log in cui le attività presentano dipendenze complesse come salti (skipping), cicli (loops) o dipendenze a lunga distanza.
- **Errata Gestione di Cicli Corti:** L'AM non riesce a identificare correttamente cicli corti (es.  $(A \rightarrow A)$  o  $(A \rightarrow B \rightarrow A)$ ). Assume erroneamente che se  $(A \rightarrow B)$  e  $(B \rightarrow A)$ , debba esistere una relazione di parallelismo  $((A \parallel B))$ , il che è sbagliato per i cicli.
- **Erronea Unione di Attività con la Stessa Etichetta:** Se la stessa etichetta di attività compare in contesti differenti all'interno del processo, l'Alpha Miner potrebbe unirle in modo scorretto, portando a modelli di processo errati.
- **Incapacità di Introdurre Transizioni Invisibili:** A differenza di algoritmi più avanzati (come l'Heuristics Miner o l'Inductive Miner), l'Alpha Miner non può introdurre transizioni invisibili (passaggi saltati) per semplificare il modello scoperto.

In sintesi, sebbene l'Alpha Miner fornisca un approccio diretto per derivare un modello di processo da un log eventi, le sue limitazioni intrinseche nella gestione di dipendenze complesse, cicli, attività con etichette duplicate e l'assenza di transizioni invisibili lo rendono meno efficace in molti scenari reali.

## Tecniche di Process Discovery: Heuristic Miner (HM) e Inductive Miner (IM)

### Contesto:

- L' $\alpha$ -algorithm, sebbene fondamentale, ha limitazioni nella gestione di log incompleti e rumorosi.
- HM e IM rappresentano evoluzioni per superare queste sfide.

### Heuristic Miner (HM)

- **Filosofia:** Filtrare il comportamento infrequente basandosi sulle frequenze di eventi e sequenze.
- **Principio Fondamentale:** Percorsi rari non dovrebbero essere inclusi nel modello.
- **Analisi:** Frequenza con cui un'attività segue direttamente un'altra.
- **Misure di Dipendenza:** Basate sulle frequenze per inferire l'ordine tra attività.
- **Vantaggi:**
  - Maggiore robustezza al rumore e all'incompletezza rispetto all' $\alpha$ -algorithm.
  - Produce modelli più compatti e comprensibili ("lasagna models" vs. "spaghetti models").
- **Output:** Rappresentazione simile ai "causal nets", capace di scoprire split e join.
- **Implementazione:** Disponibile in strumenti come ProM.
- **Limitazioni:** Può produrre modelli (C-nets) con deadlock.

### Inductive Miner (IM)

- **Filosofia:** Approccio "divide-and-conquer" (divide-et-impera), dividendo ricorsivamente il log in sub-log.
- **Passaggi Fondamentali:**
  1. **Costruzione DFG:** Simile alla relazione " $>L$ " dell' $\alpha$ -algorithm.

2. **Identificazione "Tagli" (Cuts):** Basati sugli operatori dei process tree (exclusive-choice, sequence, parallel, redo-loop) con priorità.
  3. **Divisione del Log:** In sub-log in base al tipo di taglio.
  4. **Ricorsione:** Ripetizione su ogni sub-log.
  5. **Caso Base:** Sub-log con una sola attività.
  6. **Costruzione Process Tree:** Sequenza di operatori e attività.
  7. **Soundness per Costruzione:** Il modello non soffre di deadlock.
  8. **Convertibilità:** Facile conversione in Petri net, BPMN, ecc.
- **Vantaggi:**
    1. Robusto e ampiamente utilizzato per log complessi.
    2. Modelli sempre "sound" per costruzione.
    3. Fitness perfetta (versione base).
  - **Limitazioni (Versione Base):**
    1. **Alta Variabilità:** Può produrre modelli sotto-fittati ("flower models").
    2. **Attività Duplicate:** Non gestite nativamente, può portare a underfitting.
    3. **Dipendenze a Lunga Distanza:** Difficoltà a catturarle correttamente senza duplicazione.
    4. **Focus sul Control-Flow:** Non incorpora nativamente vincoli temporali, risorse, dipendenze sui dati.
  - **Varianti per Superare le Limitazioni:**
    1. **IMF (Infrequent):** Gestione del rumore filtrando comportamenti rari.
    2. **IMC (Incompleteness):** Gestione di log incompleti.
    3. **IMD (Directly-Follows based):** Lavora direttamente sul DFG, scalabile per log enormi (ma non garantisce fitness perfetta).

#### In Sintesi:

Sia HM che IM rappresentano significativi progressi rispetto all' $\alpha$ -algorithm. HM si concentra sulla semplificazione tramite il filtraggio della frequenza. IM utilizza una strategia strutturata per garantire modelli validi e scalabili, con varianti che affrontano specifici problemi come rumore e incompletezza.

#### Genetic Miner: Process Discovery con Algoritmi Evoluzionistici

- **Approccio Iterativo:** Simula l'evoluzione naturale per scoprire modelli di processo.
- **Popolazione Iniziale:** Insieme di modelli di processo ("individui"), generati casualmente o con euristiche basate sulle frequenze del log.
- **Obiettivo (Goal):** Ottimizzare i modelli per riflettere al meglio il log eventi.
- **Funzione di Fitness:** Valuta la qualità di ogni modello bilanciando fitness, semplicità, precisione e generalizzazione. Penalizza modelli imprecisi e incompleti.
- **Genotipo:** Codifica del modello per l'ottimizzazione, tipicamente una matrice causale (che può essere interpretata come una variante di C-nets). Definisce il bias rappresentazionale.
- **Evoluzione:** Basata su operazioni genetiche:
  - **Mutazione:** Modifica casuale di un singolo individuo (es. aggiunta/rimozione di attività negli insiemi di input/output).
  - **Crossover:** Combina informazioni da due genitori per creare figli (es. scambio di parti degli insiemi di input/output).
  - **Riparazione:** Azioni post-operazioni genetiche per garantire la validità del modello (es. WF-net o C-net validi).
- **Selezione:** Individui scelti per le operazioni genetiche in base alla loro fitness (es. elitismo, selezione a torneo).
- **Arresto:** Criteri come raggiungimento di un certo livello di fitness, numero massimo di iterazioni, o mancanza di miglioramenti significativi.

- **Output:** Il modello con la fitness migliore nella popolazione finale.

#### **Vantaggi:**

- **Flessibilità e Robustezza:** Gestisce rumore e incompletezza nei log.
- **Scoperta della Concorrenza:** Capacità intrinseca.
- **Adattabilità della Fitness:** Possibilità di privilegiare specifiche caratteristiche del modello.
- **Parallelizzabilità:** L'approccio si presta all'esecuzione parallela.

#### **Limitazioni:**

- **Efficienza per Log e Modelli Grandi:** Potenziale elevato tempo di calcolo.
- **Garanzia di Validità del Modello:** A differenza dell'Inductive Miner (process tree), può scoprire modelli con deadlock (es. C-nets).
- **Log ad Alta Variabilità ("Spaghetti processes"):** Possibile output di modelli complessi ("spaghetti models").

#### **Miglioramenti:**

- **Combinazione con Euristiche:** Per la generazione della popolazione iniziale o altre fasi.

#### **In Sintesi:**

Il Genetic Miner offre un approccio di ottimizzazione iterativa per la Process Discovery, efficace nella gestione di log non perfetti grazie ai principi evolutivisti. Tuttavia, l'efficienza e la garanzia di validità del modello possono rappresentare delle sfide rispetto ad altre tecniche.

## **Validazione dei Modelli Predittivi**

#### **Introduzione:**

- Nella Data Mining, l'**accuratezza** di un modello predittivo è fondamentale.
- La **suddivisione dei dati** è un approccio standard per la selezione e la valutazione dei modelli.

#### **Suddivisione dei Dati:**

- Metodo generale: **training sample, validation sample, test sample**.
- **Cross-validation:** Partizionamento casuale del dataset in  $k$  sottoinsiemi uguali.
  - Un sottoinsieme per la **validation/test**.
  - I restanti  $k-1$  per il **training**.
  - Suddivisioni comuni: 70% training / 30% test; 50% training / 25% validation / 25% test.
- **Training:** Utilizzo di variabili di input (X) per predire una variabile target (Y).
- **Validazione/Test:** Verifica dell'accuratezza delle predizioni.

#### **Valutazione dei Classificatori: Matrice di Confusione:**

- Riassume i risultati confrontando classi attuali e predette.
- Componenti per problemi binari:
  - **True Positive (TP):** Positivi attuali classificati correttamente come positivi.
  - **False Positive (FP):** Negativi attuali classificati erroneamente come positivi (tasso di falso allarme).
  - **True Negative (TN):** Negativi attuali classificati correttamente come negativi.
  - **False Negative (FN):** Positivi attuali classificati erroneamente come negativi.

#### **Metriche di Performance (basate sulla matrice di confusione):**



- **Precision:**  $TP / (TP + FP)$  - Proporzione di istanze classificate come positive che sono effettivamente positive.
- **Recall (True Positive Rate):**  $TP / (TP + FN)$  - Proporzione di istanze effettivamente positive che sono state correttamente classificate come positive.

#### Importanza della Valutazione su Dati Non Visti:

- Le metriche calcolate solo sui dati di training sono insufficienti e statiche.
- I modelli possono ottenere punteggi perfetti sui dati di training senza generalizzare bene.
- È cruciale valutare il modello su **dati di test/validazione** differenti dal training per stimare la **generalità**.
- Richiede la "**ground truth**" nei dati di test.

#### Overfitting e Underfitting:

- **Underfitting (Under-specified):**
  - Complessità del modello troppo bassa.
  - Basse Precision e/o Recall sui dati di test.
  - Modello troppo generale, manca di precisione.
- **Overfitting (Over-specified):**
  - Modello troppo legato ai dati di training.
  - Alta accuratezza sui dati di training, bassa sui nuovi dati.
  - Eccessiva complessità, scarsa generalizzazione.

#### Fattori che Influenzano Overfitting e Underfitting:

- **Distorsione dei dati (Bias):** Dati non rappresentativi del dominio reale ("cherry-picking"). La scelta della classe di modelli (il bias) influenza l'errore di test.
- **Variabilità del dominio (Variance):** Incapacità del modello di catturare la variabilità intrinseca dei dati (dati di training come campione limitato).

#### Trade-off Bias-Variance:

- Bilanciare overfitting e underfitting è fondamentale per la predizione su nuovi dati.
- Modello a basso bias (complesso) = alta varianza (potenziale overfitting).
- Modello ad alto bias (semplice) = bassa varianza (potenziale underfitting).

#### Process Mining e Valutazione dei Modelli Scoperti:

- Nel Process Mining, si valutano modelli di processo scoperti da event log.
- Obiettivo: modello che spieghi il log e sia utile per audit, ottimizzazione, implementazione.
- **Quattro Dimensioni di Qualità:**
  1. **Fitness (Recall):** Il modello dovrebbe consentire il comportamento visto nel log (evitare "non-fitting").
  2. **Precision:** Il modello non dovrebbe consentire comportamenti estranei al log (evitare "underfitting").
  3. **Generalization:** Il modello dovrebbe generalizzare il comportamento d'esempio (evitare "overfitting").
  4. **Simplicity:** Il modello non dovrebbe essere inutilmente complesso (leggibilità).

#### Fitness (Recall) nel Process Mining:

- Misura quanto il comportamento del log è riflesso nel modello.
- Un modello con buona fitness può "riprodurre" la maggior parte delle tracce.
- Correlato al Recall nella classificazione.
- Valutazione tramite **token-based replay**: si contano token consumati (c), prodotti (p), mancanti (m), e rimanenti (r) durante la riproduzione del log sul modello (Petri net).

- **Formula Fitness:**  $f = 21 \left( 1 - \sum_{L_i \in L_{cN, \sigma}} L_i \right) + 21 \left( 1 - \sum_{L_i \in L_{pN, \sigma}} L_i \right) + 21 \left( 1 - \sum_{L_i \in L_{rN, \sigma}} L_i \right)$
- Valore tra 0 (scarsa fitness) e 1 (fitness perfetta).
- L'Inductive Miner (IM) base garantisce fitness perfetta.

#### **Precision nel Process Mining:**

- Il comportamento del modello dovrebbe riflettere accuratamente il log (evitare "underfitting").
- Il modello non dovrebbe permettere comportamenti non presenti nel log.
- Valutazione confrontando le attività possibili nel modello in un certo stato con quelle effettivamente osservate nel log con la stessa storia.
- **Formula Precision (concettuale):** Media su tutti gli eventi del rapporto tra attività osservate e permesse.
- Legata al concetto di "escaping edges" (comportamenti permessi dal modello ma non nel log).
- Valore tra 0 e 1 (1 = alta precisione).

#### **Generalization nel Process Mining:**

- Il modello dovrebbe generalizzare il comportamento d'esempio (evitare "overfitting").
- Misurazione difficile: stima di quanto bene il modello descrive un sistema sconosciuto.
- Metodo pratico: frequenza di visita dei sottoinsiemi del modello durante il replay del log.
- Buona generalizzazione = tutti i sottoinsiemi del modello visitati frequentemente.
- Valutazione tramite **cross-validation**: fitness del modello appreso su una parte del log di test.
- Trade-off con la precisione.

#### **Simplicity nel Process Mining:**

- Il modello non dovrebbe essere inutilmente complesso (Rasoio di Occam).
- Migliora la comprensione umana.
- Valutata in termini di complessità strutturale (dimensione del modello, "strutturatezza").
- Misurazione tramite la media ponderata del grado di posti/transizioni in una Petri net.
- Trade-off con fitness e precisione.

#### **Bilanciamento delle Metriche:**

- Le quattro metriche sono spesso in competizione.
- Non è facile eccellere in tutte contemporaneamente.
- Trade-off tra underfitting (bassa precisione) e overfitting (bassa generalizzazione).
- Trade-off tra fitness e semplicità.
- Algoritmi diversi gestiscono questi trade-off in modo differente.

#### **Influenza della Qualità del Log Eventi:**

- Log incompleti e rumorosi rappresentano una sfida.
- Rumore: comportamenti rari e non tipici.
- Incompletezza: log insufficiente per scoprire la struttura.
- Algoritmi avanzati (heuristic, genetic, inductive mining) cercano di gestire questi problemi.
- La qualità del modello può essere migliorata filtrando il rumore o usando l'analisi delle varianti.

#### **Conclusione:**

La valutazione dei modelli di processo scoperti è multidimensionale e complessa. Le quattro metriche (Fitness, Precisione, Generalizzazione e Semplicità) sono essenziali e devono essere

bilanciate per ottenere modelli utili e affidabili. La qualità del log eventi è un fattore cruciale che influenza la qualità del modello scoperto.

## Analisi della Qualità del Modello di Processo Scoperto

### Scenario di Qualità del Modello di Processo Scoperto:

- **Fitness (Recall) alta:** Il modello riproduce bene il comportamento osservato nel log di training.
- **Precisione alta:** Il modello è rigoroso e permette solo comportamenti strettamente osservati nel log (evita l'underfitting).
- **Generalizzazione bassa:** Il modello è troppo specifico per il log di training e potrebbe non gestire bene dati non visti (sintomo di overfitting).
- **Semplicità bassa:** La struttura del modello è complessa.
- **Comprendibilità alta:** (Potenzialmente inusuale con bassa semplicità, potrebbe riferirsi a chiarezza delle etichette o altri aspetti non strutturali).

### Interpretazione:

- Modello con **overfitting**: si adatta troppo specificamente al log di training, diventando complesso e non generalizzando bene a nuovi dati.
- **Trade-off** tra le quattro dimensioni di qualità (fitness, precisione, generalizzazione, semplicità).

### Valutazione degli Algoritmi di Discovery:

- Gli algoritmi vengono confrontati sperimentalmente su metriche di qualità utilizzando diversi event log.
- Si costruiscono classifiche degli algoritmi per ciascuna metrica e una classifica finale media.
- La relazione tra algoritmi e metriche dipende dalla variabilità del log.
- La valutazione è complessa a causa delle molteplici dimensioni e rappresentazioni.
- La letteratura fornisce approcci per quantificare le quattro dimensioni di qualità.

### Miglioramento della Qualità del Modello:

- **Riduzione della variabilità** nell'event log di input:
  - **Filtrare noise:** Comportamento irrilevante o poco frequente ("outliers").
  - **Analisi delle varianti:** Segmentare il log per spiegare le differenze tra i casi.
  - **Uso di regole:** Segmentare il log basandosi su caratteristiche specifiche dei casi.
  - **Clustering dell'event log:** Raggruppare tracce simili in sottolog più omogenei.
- Confronto della qualità del modello per segmenti dell'Event Log.
- Applicazione iterativa di filtraggio o segmentazione per ottenere modelli di alta qualità.

### Scopi del Process Discovery:

- Derivare il modello di processo sottostante da un event log.
- Si posiziona tra Data Mining e Business Process Management (BPM).
- Aiuta le organizzazioni a scoprire i loro processi attuali.
- Correlato agli obiettivi di comprensione, descrizione e previsione del business.
- **Scopi Specifici:**
  - Comprensione del business (identificazione e analisi dei processi).
  - Descrizione del business (reporting, segmentazione, identificazione di comportamenti).
  - Previsione del business (stima, classificazione).
  - Ottimizzazione dei processi (identificazione di colli di bottiglia e costi).
  - Automazione dei processi (trovare soluzioni più veloci ed economiche).

- Conformance Validation (Checking) (verifica della conformità del log al modello).
- Simulazione dei processi (previsioni future).
- Enhancement dei processi (estensione o miglioramento di modelli esistenti).
- È un compito impegnativo con un complesso spazio di ricerca.

**Elemento Mancante (da considerare):**

- **Trade-off specifici tra le metriche di qualità:** Potrebbe essere utile menzionare più esplicitamente i trade-off comuni, ad esempio come un tentativo di aumentare la precisione possa diminuire la generalizzazione, o come un modello più semplice possa avere una fitness inferiore. Questo rafforzerebbe la comprensione della sfida nel trovare un modello "ottimale".

# Conformance Checking

## Conformance Checking nel Process Mining (M3-L4)

Il **Conformance Checking** è una delle tre aree principali del Process Mining, insieme alla **Process Discovery** (scoperta di processi) e alla **Process Enhancement** (miglioramento di processi).

### 1. Obiettivo e Definizione del Conformance Checking

L'obiettivo fondamentale del Conformance Checking è **identificare e misurare la gravità delle deviazioni** tra l'esecuzione effettiva di un processo aziendale, registrata nel **log eventi**, e un insieme di **specifiche prescrittive**.

In altre parole, il Conformance Checking confronta:

- **Processo Osservato (Observed Process):** Ciò che è accaduto in realtà, come documentato negli *event log*. Un log eventi è intrinsecamente finito e contiene solo un comportamento di esempio, che può coprire solo parzialmente il comportamento completo del processo. La distribuzione delle tracce è anch'essa importante e può essere tenuta in considerazione tramite funzioni di probabilità.
- **Processo Atteso (Expected Process):** Come le cose dovrebbero accadere, definito da modelli di processo (imperativi o dichiarativi) o da regole di business. Un modello di processo, invece, può generare un numero potenzialmente infinito di tracce (specialmente con iterazioni ed esecuzioni concorrenti) e fornisce una vista globale del processo.

Questo confronto mira a trovare sia **punti in comune** che **discrepanze**.

*(Qui sarebbe posizionata l'immagine "Image 1" da 3.5-ConformanceChecking.pdf, pag. 2. Rappresenta schematicamente il Conformance Checking, mostrando il confronto tra il processo osservato e quello atteso, spesso con un diagramma di flusso del processo su uno schermo di computer e un telefono.)*

### 2. Scopi e Applicazioni Specifiche del Conformance Checking

Il Conformance Checking è rilevante per diversi scopi specifici e trova applicazioni pratiche significative:

- **Business Alignment e Auditing:**
  - Verifica se la realtà, come registrata nel log, è conforme a un modello e viceversa. Ad esempio, può accertare se le procedure aziendali stabilite vengono effettivamente seguite.
  - L'analisi del log eventi può mostrare se una regola (es. l'obbligo di due controlli per ordini superiori a un milione di Euro) viene rispettata.
  - Può essere utilizzato per rilevare, localizzare, spiegare le deviazioni e misurarne la gravità.

- **Compliance Checking:**
  - Verifica che i processi aziendali rispettino leggi, regolamenti esterni, politiche aziendali interne e procedure.
  - Può aiutare a scoprire casi di frode scansionando il log eventi con un modello che specifica i requisiti, come il principio dei "quattro occhi" (segregazione dei compiti).
  - Le tecniche di Process Mining offrono un mezzo per un controllo di conformità più rigoroso e per accertare la validità e l'affidabilità delle informazioni sui processi.
- **Valutazione degli Algoritmi di Process Discovery:**
  - Le tecniche di Conformance Checking sono utilizzate per valutare la qualità dei modelli scoperti dagli algoritmi di Process Discovery. Un buon algoritmo di discovery dovrebbe produrre modelli con alta fitness e precisione rispetto al log da cui sono stati estratti.
- **Riparazione di Modelli (Model Repair):**
  - Le discrepanze identificate possono suggerire adattamenti del modello per farlo riflettere meglio la realtà osservata nel log.
- **Diagnosi delle Deviazioni:**
  - Fornisce una diagnostica dettagliata su dove e come avvengono le deviazioni.
- **Supporto Operazionale:**
  - Può essere utilizzato in un contesto online per rilevare deviazioni in tempo reale, mentre i processi sono ancora in esecuzione ("on-the-fly"). Questo consente interventi tempestivi per correggere il tiro o mitigare i rischi.

L'interpretazione della **non conformità** dipende criticamente dallo **scopo del modello**:

- Se il modello è **descrittivo**, le discrepanze indicano che il modello deve essere migliorato per catturare meglio la realtà.
- Se il modello è **normativo** (prescrittivo), le discrepanze possono:
  - Esporre **deviazioni indesiderabili** (segnalando la necessità di un migliore controllo del processo).
  - Rivelare **deviazioni desiderabili** (dove i lavoratori deviano dalla procedura standard per servire meglio i clienti o gestire circostanze impreviste, suggerendo potenziali miglioramenti al modello normativo stesso).

### 3. Approcci al Conformance Checking

Le fonti identificano principalmente due approcci al Conformance Checking:

#### 3.1. Model-based Conformance Checking

- **Concetto:** Implica un **Modello di Processo Imperativo**, come una Rete di Petri o un modello BPMN. L'obiettivo è verificare se le tracce nel log eventi possono essere "eseguite" o "riprodotte" (**replay**) seguendo i passi consentiti dal modello.
- **Comportamento del Modello:** Gli eventi coinvolti in iterazioni ed esecuzioni concorrenti portano un modello di processo a generare un comportamento potenzialmente infinito.
- **Tecniche chiave:**
  - **Token Replay:** Una tecnica in cui le tracce del log vengono riprodotte su una Rete di Petri. Si contano i "token mancanti" (quando un'attività nel log si verifica ma non è abilitata nel modello) e i "token rimanenti" (quando un token rimane nel modello dopo la riproduzione completa della traccia). Viene calcolata una metrica di fitness (idoneità) basata su questi conteggi (es. tra 0 e 1).
    - **Limiti:** Può fornire diagnostiche fuorvianti a causa dei token rimanenti e i valori di fitness possono essere generalmente troppo bassi. Assume l'uso di una Rete di Petri.
  - *(Qui sarebbe posizionata l'immagine "Image 6" da 3.5–ConformanceChecking.pdf, pag. 14. Mostra un modello di processo semplice con un ciclo, illustrando come possa generare comportamenti infiniti come ABCABCABCE, ABCE, ABCABCE.)*  
*(Qui sarebbe posizionata l'immagine "Image 7" da 3.5–ConformanceChecking.pdf, pag. 15. Un altro esempio di modello di processo con percorsi concorrenti e ciclici, che può generare comportamenti come ABCCBCBE, ABCCCCBCBE, ma non ABCABCE.)*  
*(Qui sarebbe posizionata l'immagine "Image 10" da 3.5–ConformanceChecking.pdf, pag. 19. Illustrazione del replay di una traccia su una Rete di Petri, mostrando i token e le transizioni. L'immagine mostra una sequenza di stati di una Petri net durante il replay di una traccia.)*  
*(Qui sarebbe posizionata l'immagine "Image 11" da 3.5–ConformanceChecking.pdf, pag. 20. Esempio più dettagliato di replay con l'indicazione dei token mancanti e rimanenti e le metriche associate. Mostra un modello di Petri net e diverse tabelle che riassumono i risultati del replay, inclusi i token consumati, prodotti, mancanti e rimanenti per diverse tracce.)*  
*(Qui sarebbe posizionata l'immagine "Image 12" da 3.5–ConformanceChecking.pdf, pag. 21. Ulteriore esempio di replay, mostrando le deviazioni e il calcolo della fitness. Simile all'immagine precedente, ma con un focus su diverse tracce e i relativi risultati di conformità.)*
  - **Alignments (Allineamenti):** Questa tecnica stabilisce un accoppiamento stretto tra il log eventi e il modello di processo. Per ogni traccia nel log, identifica il percorso "più vicino" nel modello che potrebbe averla generata, quantificando i "costi" di disallineamento (chiamati "moves"):
    - **Mosse Sincrone:** Log e modello eseguono la stessa attività contemporaneamente (es.  $(a, (a, t1))$ ). Generalmente non hanno costi.

- **Mosse del Modello (Visibili):** Il modello esegue un'attività visibile (con etichetta diversa da  $\tau$ ) che non è presente nel log in quel punto (es.  $(\gg, (y, \tau))$ ). Hanno un costo (tipicamente 1). Sono anche chiamati "mismatches" ottenuti introducendo un simbolo di salto (-) nel modello (Classificati come **Skip**).
  - **Mosse del Modello (Invisibili):** Il modello esegue una transizione silenziosa ( $\tau$ ) che non ha una corrispondente attività nel log (es.  $(\gg, (\tau, \tau))$ ). Considerate innocue, generalmente hanno costo 0.
  - **Mosse del Log:** Il log registra un evento di un'attività (x) che non ha una corrispondente mossa nel modello in quel punto (es.  $(x, \gg)$ ). Hanno un costo (tipicamente 1). Sono anche chiamati "mismatches" ottenuti introducendo un simbolo di salto (-) nel log (Classificati come **Insert**).
  - **Obiettivo:** Trovare un allineamento che minimizzi il costo totale delle mosse non sincrone. I costi possono essere configurati. L'allineamento ottimale non è necessariamente unico.
  - **Vantaggi:** Permettono di mappare il comportamento osservato su quello modellato anche in caso di non conformità, fornendo diagnostiche più dettagliate e accurate rispetto al Token Replay. I costi di disallineamento possono essere convertiti in un valore di fitness tra 0 e 1 (calcolato come  $1 - (\text{costo allineamento ottimale} / \text{costo allineamento peggiore})$ ). Sono indipendenti dal modello (funzionano con qualsiasi notazione con semantica formale) e sono configurabili tramite una funzione di costo.
  - **Svantaggi:** Il costo computazionale è elevato, può essere esponenziale rispetto alla lunghezza delle tracce. Le tecniche di allineamento delle tracce non gestiscono esplicitamente i task concorrenti né il comportamento ciclico, perché queste caratteristiche non sono direttamente osservabili a livello delle singole tracce.
  - **Origine:** L'algoritmo di allineamento delle tracce ha origine dalla bioinformatica, dove è utilizzato per allineare sequenze proteiche e geniche al fine di identificare strutture comuni e mutazioni.
  - **Allineamento di più tracce:** Per allineare più tracce, un'idea di base è allineare progressivamente le tracce a coppie. La selezione delle tracce per l'allineamento ad ogni iterazione si basa sulla loro similarità. Le tracce più simili tra loro vengono allineate per prime. Una volta che tracce simili sono state allineate, si allineano i cluster di tracce risultanti l'uno contro l'altro. Un "albero guida" (guide tree) viene costruito per assistere questo processo, tipicamente utilizzando algoritmi di clustering gerarchico agglomerativo (AHC).
- *(Qui sarebbe posizionata l'immagine "Image 13" da 3.5–ConformanceChecking.pdf, pag. 24. Rappresentazione schematica di un allineamento di tracce, mostrando le mosse sincrone e asincrone tra log e modello. L'immagine mostra una visualizzazione a matrice o a griglia con diverse tracce allineate, evidenziando le attività comuni e le deviazioni.)*



(Qui sarebbe posizionata l'immagine "Image 14" da 3.5-  
*ConformanceChecking.pdf*, pag. 25. Esempio di allineamento con le  
formule per il calcolo della fitness basata sui costi. L'immagine mostra tre esempi di  
allineamenti tra tracce e modelli, con le relative formule di fitness che considerano i  
"mismatches".)

- **Metriche di Qualità del Modello (anche dopo la Discovery, ma rilevanti per la Conformità):**
  - **Fitness (Recall):** Misura la capacità del modello di riprodurre il comportamento osservato nel log. Formula:  $\text{Fitness}(M,L)=LM \cap L$ . Un valore tra 0 (idoneità molto scarsa) e 1 (idoneità perfetta). C'è una buona convergenza su questa metrica.
  - **Precisione:** Il modello non dovrebbe permettere comportamenti non visti nel log ("evitare l'underfitting"). Misura quanto del comportamento consentito dal modello è effettivamente osservato nel log. Formula:  $\text{Precision}(M,L)=MM \cap L$ . Poiché il comportamento di un modello può essere infinito, esistono approcci concorrenti in letteratura per la Precision.
  - **Generalizzazione:** Il modello dovrebbe generalizzare il comportamento visto nel log senza essere troppo specifico ("evitare l'overfitting").
  - **Semplicità:** Il modello non dovrebbe essere inutilmente complesso (Rasoio di Occam).

### 3.2. Rule-based Conformance Checking

- **Concetto:** Identifica **vincoli o regole (Business Rules)** a cui il comportamento eseguito deve rispettare. Questo approccio si basa su un **Modello di Processo Dichiarativo** se il comportamento è definito usando la logica temporale, altrimenti può riferirsi a un modello imperativo.
- **Regole come vista locale:** Le regole forniscono una vista locale del processo, filtrando le tracce non conformi.
- **Dimensioni vincolate:** Le regole possono riguardare:
  - **Control-flow:** Ordine di esecuzione delle attività.
  - **Time:** Vincoli temporali tra attività o scadenze.
  - **Resource:** Vincoli su chi può/non può eseguire determinate attività (es. principio dei "quattro occhi").
  - **Other attributes:** Vincoli su altri attributi degli eventi o dei casi.
- **Complessità delle Regole:**
  - **Semplici:** Possono essere definite tramite espressioni regolari, usando ad esempio un Automa a Stati Finiti (Finite State Automaton).

- **Avanzate:** Richiedono l'uso della **Logica Temporale** per verificare condizioni pre e post locali e non locali. Un esempio è la Linear Temporal Logic (LTL), che utilizza operatori temporali come "sempre" ( $\Box$ ), "eventualmente" ( $\Diamond$ ), "fino a" (U). La logica temporale può includere anche riferimenti espliciti al tempo e ai dati.
- **Modelli Imperativi vs. Dichiarativi:**
  - **Modelli Imperativi:** (es. Petri Nets, BPMN) Rappresentano l'intero comportamento del processo in una volta sola. Sono appropriati per domini in cui può essere imposto un controllo centrale sul modello.
  - **Modelli Dichiarativi:** Si basano su un insieme di regole che vincolano il comportamento. Sono appropriati per descrivere ambienti dinamici, dove i processi sono altamente flessibili e soggetti a cambiamenti. La filosofia è: "tutto è possibile a meno che non sia esplicitamente vietato".
- **DECLARE:** È un linguaggio di modellazione di processi dichiarativi basato sui vincoli e con semantica basata su LTL. Include diverse costrizioni comuni con formalizzazione LTL, come:
  - **Existence(A):** L'attività A si verifica almeno 1 volta.
  - **Absence(A):** L'attività A non si verifica.
  - **Init(A):** L'attività A è la prima a verificarsi.
  - **Last(A):** L'attività A è l'ultima a verificarsi.
  - **Choice(A,B):** A o B si verificano.
  - **Exclusive Choice(A,B):** A o B si verificano, ma non insieme.
  - **RespondedExistence(A, B):** Se A si verifica, allora B si verifica.
  - **Co-Existence(A, B):** Se A si verifica allora B si verifica, e se B si verifica allora A si verifica.
  - **Response(A, B):** Se A si verifica, allora B si verifica dopo A.
  - **AlternateResponse(A, B):** Ogni volta che A si verifica, B si verifica successivamente, prima che A si ripeta.
  - **ChainResponse(A, B):** Ogni volta che A si verifica, B si verifica immediatamente dopo.
- *(Qui sarebbe posizionata l'immagine "Image 2" da 3.5–ConformanceChecking.pdf, pag. 7. Mostra una tabella con esempi di vincoli DECLARE come Existence, Absence, Init, Last, con la loro formalizzazione LTL e alcuni esempi di tracce che li soddisfano o li violano.)*
- *(Qui sarebbe posizionata l'immagine "Image 3" da 3.5–ConformanceChecking.pdf, pag. 8. Mostra una tabella con esempi di vincoli DECLARE come Choice ed Exclusive Choice, con la loro formalizzazione LTL e alcuni*

*esempi di tracce.)*

*(Qui sarebbe posizionata l'immagine "Image 4" da 3.5–*

*ConformanceChecking.pdf, pag. 9. Mostra una tabella con esempi di vincoli DECLARE come RespondedExistence e Co-Existence, con la loro formalizzazione LTL e alcuni esempi di tracce.)*

*(Qui sarebbe posizionata l'immagine "Image 5" da 3.5–*

*ConformanceChecking.pdf, pag. 10. Mostra una tabella con esempi di vincoli DECLARE come Response, AlternateResponse e ChainResponse, con la loro formalizzazione LTL e alcuni esempi di tracce.)*

- **Librerie per la modellazione dichiarativa:**
  - **RuM:** ([rulemining.org](http://rulemining.org))
  - **PM4PY:** ([pm4py.fit.fraunhofer.de](http://pm4py.fit.fraunhofer.de))
- **Scopi del Rule-based Conformance Checking:**
  - Verifica di Conformità e Audit: Controllo diretto del rispetto di leggi, regolamenti, politiche aziendali e procedure.
  - Identificazione di Deviazioni: Le regole possono filtrare le tracce non conformi o dividere il log in casi conformi/non conformi.
  - Analisi delle Cause Fondamentali (Root-Cause Analysis): Il filtraggio basato su regole aiuta a capire perché si verificano determinate deviazioni.
  - Monitoraggio in Tempo Reale: I vincoli di conformità possono essere verificati "al volo" sui processi in esecuzione.
  - Valutazione (implicita): L'applicazione di regole fornisce una misura di allineamento a un vincolo specifico.
- **Sfide:** Acquisizione, modellazione e formalizzazione dei vincoli rilevanti; gestione delle violazioni; interazione con gli utenti.

### 3.3. Exceptional Analysis (Analisi delle Eccezioni)

- Correlata al Conformance Checking, si concentra sull'analisi dei casi che deviano dai percorsi standard, cercando di comprendere le eccezioni piuttosto che solo misurare la conformità.

### 3.4. Comparing Footprints (Confronto di Impronte)

- **Concetto:** Confronta matrici (footprints) che caratterizzano il log eventi e il modello di processo sulla base di relazioni d'ordine tra attività (come la relazione "directly-follows").
- **Misura:** Una maggiore differenza tra le matrici indica minore conformità.
- **Vantaggi:** Più semplice da implementare.
- **Svantaggi:** Fornisce meno dettagli sugli specifici punti di deviazione e richiede che il log sia "completo" rispetto alla relazione d'ordine considerata.

## 4. Conformance Checking Purposes (Riepilogo)

Il Conformance Checking persegue scopi fondamentali e interconnessi:

- **Identificare il comportamento deviante nel Log Eventi:** Scoprire il comportamento osservato nel log che non è consentito dal modello (spesso definito "comportamento disfunzionale").
- **Identificare comportamenti aggiuntivi per aggiornare il Modello:** Rilevare comportamenti che sono consentiti dal modello ma che non sono mai stati osservati nel Log Eventi (definiti "comportamento non specificato"). Questo può indicare un modello troppo permissivo o la presenza di percorsi nel processo che non vengono mai attivati nella realtà.
- **Combinazione dei due scopi:** L'obiettivo tipico è duplice: da un lato, rimuovere gli errori banali (le deviazioni) e, dall'altro, apprendere nuovi comportamenti da specificare, al fine di migliorare sia il modello che la comprensione del processo stesso.
- **Confronto tra segmenti:** Il Conformance Checking può essere sfruttato per confrontare il comportamento di due segmenti del processo (ad esempio, diverse regioni geografiche, periodi temporali distinti o gruppi specifici di risorse). Questo permette di verificare se il numero e la natura delle deviazioni differiscono tra i segmenti, supportando analisi comparative e attività di benchmarking.

In sintesi, il Conformance Checking mira a **rimuovere gli errori banali nell'esecuzione del log** e **apprendere nuovi comportamenti da specificare**, collegando **Conoscenza Descrittiva** (log eventi, regole di business, modelli di processo) e **Conoscenza Prescrittiva** (violazioni) per arrivare a **Regole di Business scoperte** e **Modelli di Processo scoperti**.

*(Qui sarebbe posizionata l'immagine "Image 94" da 3.5-ConformanceChecking.pdf, pag. 36. Un diagramma che illustra come il Conformance Checking colleghi la Conoscenza Descrittiva (Event Log, Business Rules, Business Process Models) alla Conoscenza Prescrittiva (Violations) per generare nuove Regole di Business e Modelli di Processo scoperti. Mostra frecce che collegano Event Log, Business Rules e Business Process Models a "Violations" tramite "Conformance Checking", portando a "Discovered Business Rules" e "Discovered Process Models".)*

## 5. Classificazione delle Attività Disallineate (con Trace Alignment)

Dall'analisi delle "mosse" non sincrone prodotte dagli algoritmi di allineamento delle tracce, è possibile classificare le attività disallineate e identificare le ragioni delle anomalie. Questa classificazione fornisce una diagnostica dettagliata e facile da capire:

### 1. Insert (Inserimento):

- **Definizione:** Attività eseguita nel log ma **non specificata** nel modello nel punto corrispondente dell'allineamento.
- **Caratteristica:** Una mossa del log dove l'attività appare nella riga del log (es. ( x , >> ) ) ma non c'è una mossa corrispondente nel modello.
- **Identificazione:** L'attività si trova nell'elenco delle mosse del log ma *non* nell'elenco delle mosse del modello per quella deviazione.

- **Significato:** Indica un'attività "extra" o non prevista nel processo reale rispetto al modello.

## 2. Skip (Salto):

- **Definizione:** Attività **specificata nel modello** ma **non eseguita** nella traccia nel punto corrispondente dell'allineamento.
- **Caratteristica:** Una mossa del modello (visibile) dove l'attività appare nel percorso del modello (es.  $(\gg, (y, t))$ ) ma non c'è una mossa corrispondente nel log.
- **Identificazione:** L'attività si trova nell'elenco delle mosse del modello ma *non* nell'elenco delle mosse del log per quella deviazione.
- **Significato:** Indica un'attività prevista dal modello che è stata "saltata" nell'esecuzione reale.

## 3. Early (In Anticipo):

- **Definizione:** Un'attività che è presente sia nelle mosse del log che nelle mosse del modello (in punti diversi dell'allineamento) è osservata nell'elenco delle **mosse del log prima** rispetto a quando appare nell'elenco delle mosse del modello.
- **Significato:** Implica che l'attività è stata eseguita prima rispetto alla sua posizione attesa nel modello, indicando un riordino o un'esecuzione prematura.

## 4. Late (In Ritardo):

- **Definizione:** Un'attività che è presente sia nelle mosse del log che nelle mosse del modello (in punti diversi dell'allineamento) è osservata nell'elenco delle **mosse del modello prima** rispetto a quando appare nell'elenco delle mosse del log.
- **Significato:** Implica che l'attività è stata eseguita dopo rispetto alla sua posizione attesa nel modello, indicando un riordino o un'esecuzione posticipata.

Queste classificazioni forniscono una diagnostica più dettagliata e accurata rispetto a metodi come il replay basato sui token, che potrebbero semplicemente indicare token mancanti o rimanenti senza specificare il tipo esatto di deviazione.

## 6. Epistemic Dilemma nel Process Mining (Secondo van der Aalst, 2018)

Il Process Mining, in particolare Process Discovery e Conformance Checking, non sono modelli predittivi ma tecniche per misurare l'appropriatezza tra due artefatti (log e modello). Questo solleva importanti questioni epistemologiche:

- **Il Log Eventi contiene un comportamento di esempio:** Non possiamo sapere se copre tutto il comportamento che il modello dovrebbe rappresentare (comportamento non ancora osservato o non enumerabile).
- **Il Modello è una rappresentazione:** Non possiamo sapere se rappresenta tutto il comportamento che il sistema ha reso possibile (comportamento non osservato, non rappresentabile o non rilevante).

Questo dilemma porta a importanti decisioni di design nei progetti di PM:

- **Quando  $L \setminus M$  (comportamento nel log non nel modello):** È area 2 (comportamento disfunzionale) o area 7 (comportamento non specificato, ma valido)? Servono euristiche per decidere.
- **Quando  $M \setminus L$  (comportamento nel modello non nel log):** È area 4 (specifica errata) o area 5 (comportamento non osservato)? Servono euristiche per decidere.

La risposta a queste domande dipende largamente dallo scenario e dal grado di orientamento al processo e di strutturazione del processo di business. Dipende da quanto estesa ci aspettiamo che siano le aree di "comportamento non nel log" o "comportamento non nel modello".

- **Processo normativo o emergente:** Quanto rigido è il processo atteso?
- **Processo stazionario o non stazionario:** Il processo cambia nel tempo?
- **Soggetto a varianza alta o bassa:** Quanto prevedibile è il comportamento del processo?

*(Qui sarebbe posizionata l'immagine "Image 90" da 3.5-ConformanceChecking.pdf, pag. 34. Un diagramma di Venn che illustra le relazioni tra il comportamento del Sistema (S), del Log Eventi (L) e del Modello (M), evidenziando le diverse aree di conformità e non conformità.)*

## **Trace Encoding**