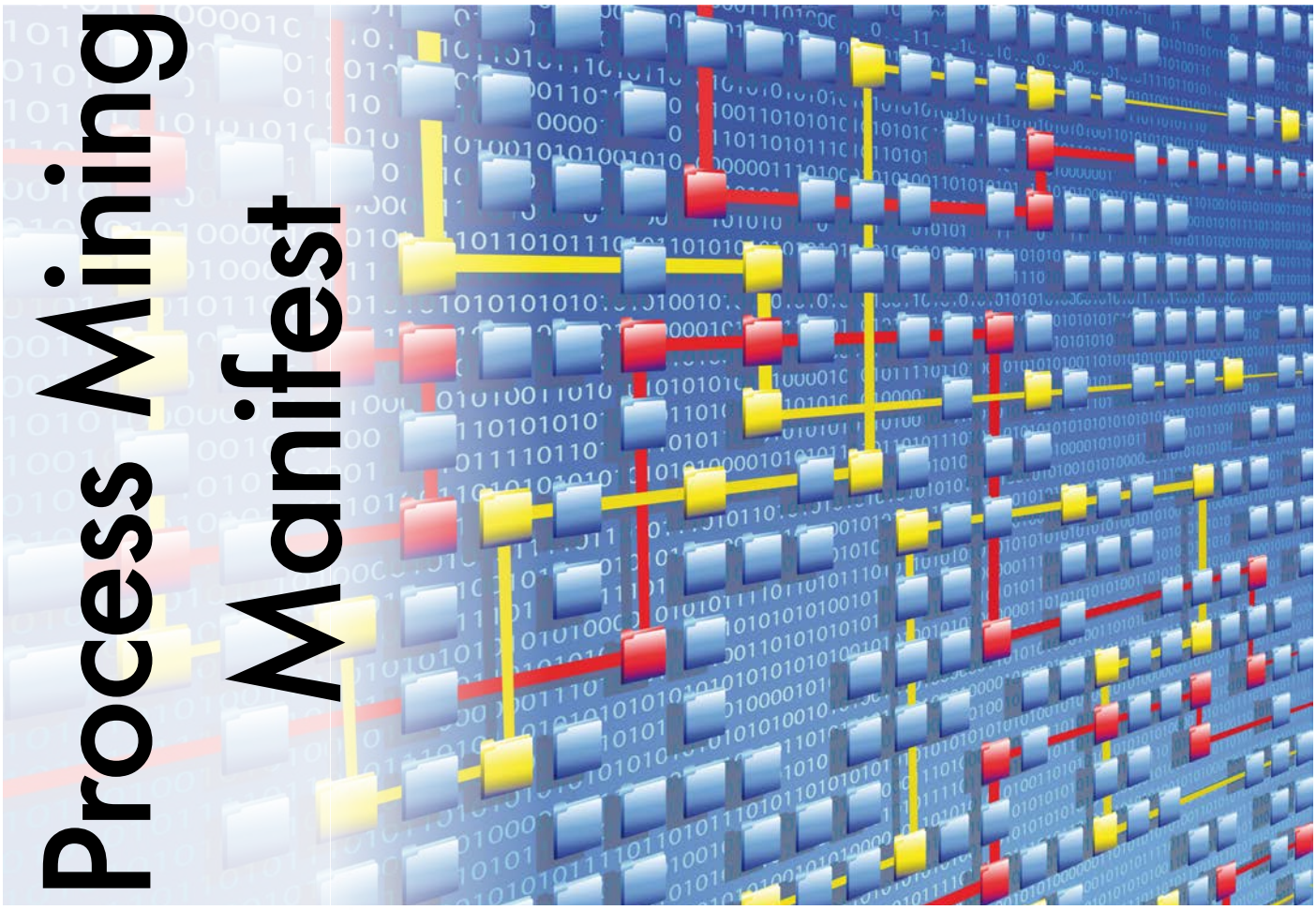
Process Mining Manifest



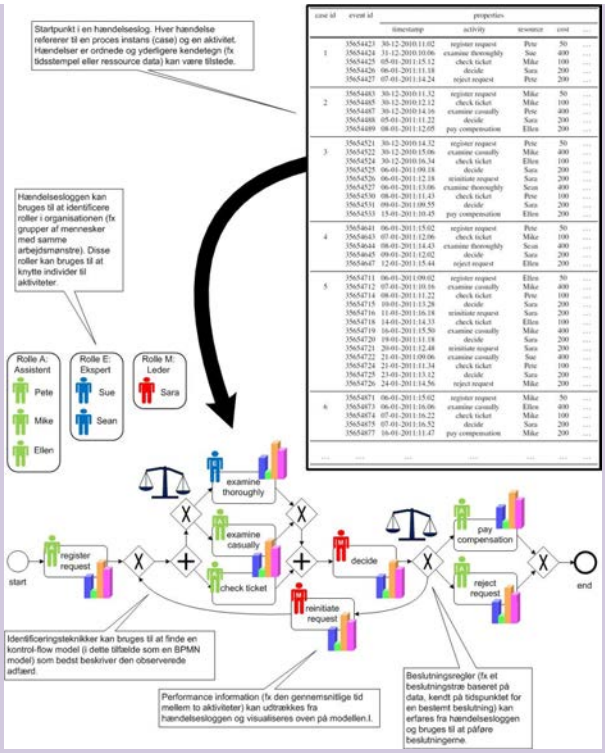
Et Manifest er en "offentligt tilgængelig erklæring vedrørende principper og intentioner" udformet af en gruppe. Dette manifest er udarbejdet af medlemmer og tilhængere af IEEE Task Force on Process Mining. Formålet med denne Task Force er at fremme research, udvikling, uddannelse, implementering og forståelse for process mining.

Process mining er en relativ ung forsknings disciplin, som befinder sig midt imellem computerstøttet intelligens og data mining på den ene side og proces modellering og analyse på den anden side. Ideen med process mining er at *identificere, monitorere og forbedre virkelige processer* (dvs. Ikke ikke-virkelige processer der bygger på antagelser) ved at trække viden ud af hændelseslogfiler, der allerede findes i vore dages (informations) systemer.

Process mining indbefatter (automatisk) proces identifikation (fx udlede en proces model fra en hændelseslog), check af overensstemmelser (for eksempel se afvigelser ved at sammenligne model og log), sociale netværk / organisatorisk mining, automatisk generering af simulations modeller, model udvidelser, model reparation, forudsigelser af forløb og historik-baserede anbefalinger.

Indhold:

Process mining teknikker kan udlede viden fra hændelseslogfiler, der ofte er tilgængelige i vore dages informations systemer. Disse teknikker giver nye metoder til at identificere, monitorere og forbedre processer i en række forskellige applikations domæner. Der er to hoved drivkræfter i den stigende interesse for process mining. På den ene side, bliver der genereret flere og flere hændelser, som giver detaljeret historik om processerne. På den anden side er der et behov for at forbedre og understøtte forretningsprocesserne i en konkurrencepræget verden under hastig forandring. Dette manifest er lavet af the IEEE Task Force on Process Mining, og har som mål at fremme emnet process mining. Derudover, ved at fastsætte et sæt vejledende principper og adressere vigtige udfordringer, er det håbet, at dette manifest kan fungere som en guide for software udviklere, forskere, konsulenter, virksomhedsledere og slutbrugere. Målet er at øge modenheden for process mining som et nyt værktøj til at forbedre, kontrollere og understøtte operationelle forretningsprocesser.



Figur1:Process mining teknikker udtrækker viden fra hændelseslogs for identificering, monitorering og forbedring af processer.

Process Mining danner en vigtig bro mellem *data mining* og *business process modelling* og *analyse*. Under Business Intelligence (BI) paraplyen er der blevet introduceret mange buzzwords, som blot henviser til forholdsvis simple rapporteringer og dashboard værktøjer. Business Activity Monitoring (BAM) henviser til teknologier, der muliggør real-tids monitorering af forretningsprocesser. Complex Event Processing (CEP) henviser til teknologier til at behandle store mængder af hændelser, bruge dem til at monitorere, stye og optimere forretningen i real-tid. Corporate Performance Management (CPM) er et andet buzzword for måling af en proces eller organisations performance. Endvidere relateret til ledelses tilgange såsom Continous Process Improvement (CPI), Business Process Improvement (BPI), Total Quality Management (TQM) og Six Sigma. Disse tilgange har det tilfælles, at processerne "lægges under mikroskop" for at blive undersøgt og se om det er muligt at forbedre processerne yderligere. Process mining er en teknologi der understøtter CPM, BPI, TQM, Six Sigma og tilsvarende teknologier.
Hvor BI-værktøjer og ledelsestilgange som Six Sigma og TQM har til formål at forbedre den operationelle performance, eksempelvis at reducere gennemløbstid og fejl, lægger organisationer også mere vægt på Corporate Governance, risiko og overholdelse af guidelines. Lovgivning såsom Sarbanes-Oxley (SOX) og Basel II illustrerer det fokus, der er på spørgsmål vedrørende overholdelse af guidelines og love. Process mining teknikker giver mulighed for, mere stringent, at kontrollere og fastslå rigtigheden og pålideligheden af information om en organisations vigtigste processer. I det seneste årti er data om hændelser blevet mere tilgængelige, og process mining teknikker er blevet modnet en del.

Kontrete mål for Task Force'n:

1) Gøre slutbrugere, udviklere, konsulenter, virksomhedsledere og forskere opmærksomme på state-of-the art i process mining
2) Fremme anvendelsen af process mining teknikker og værktøjer og stimulere til frembringelse af nye applikationer
3) Spille en rolle i standardiseringsarbejdet for at gemme (logge) hændelsesdata
4) Organisere tutorials, specielle events, workshops og panel diskussioner
5) Udgive artikler, bøger, videoer

Derudover, som før nævnt, kan ledelses trends relateret til procesforbedringer (for eksempel Six Sigma, TQM, CPI og CPM) og til overholdelse af love / guidelines drage nytte af process mining. Heldigvis er process mining algoritmer blevet implementeret i forskellige akademiske og kommercielle systemer. I dag er der en aktiv gruppe af forskere, der arbejder med process mining og det er blevet et "hot" emne indenfor Business Process Management (BPM) forskning. Endvidere, er der en stor interesse fra industrien omkring process mining. Flere og flere software leverandører tilføjer process mining funktionalitet til deres værktøjer. Eksempler hvor software produkter indeholder process mining elementer er: ARIS Process Performance Manager (Software AG), Comprehend (Open Connect), Discovery Analyst (StereoLOGIC), Flow (Fourspark), Futura Reflect (Futura Process Intelligence), Interstage Automated Process Discovery (Fujitsu), OKT Process Mining suite (Exeura), Process Discovery Focus (Iontas / Verint), ProcessAnalyzer (QPR), ProM (TU/e), Rbminer/Dbminer (UPC) og Reflect I one (Pallas Athena). Den voksende interesse i log-baseret proces analyse, har motiveret til oprettelsen af en Task Force on Process Mining.
Denne Task Force blev etableret i 2009 under Data Mining Technical Committee (DMTC) i the Institute of Electrical and Electronical Engineers (IEEE).

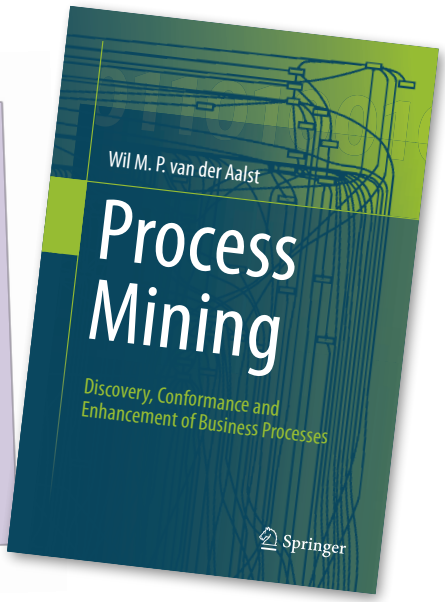
Den nuværende Task Force har medlemmer som repræsenterer software leverandører (for eksempel Pallas Athena, Software AG, Futura Process Intelligence, HP, IBM, Infosys, Fluxicon, Businesscape, Iontas/Verint, Fijitsu, Fujitsu Laboratories, Business Process Mining og Stereologic), konsulentfirmaer og slutbrugere (for eksempel ProcessGold, Business Process Trends, Gartner, Deloitte, Process Sphere, Siav SpA, BPM Chile, BWI Systeme GmbH, Excellentia BPM, Rabobank) og forskningsinstitutter (for eksempel TU/e, University of Padua, Universitat Politecnica de Catalunya, New Mexico State University, IST – Technical University of Lisbon, University of Calabria, Penn State University, University of Bari, Humbolt-Universität zu Berlin, Queensland University, of Technology, Vienna University of Economics and Business, Stevens Institute of Technology, Universita of Haifa, University of Bologna, Ulsan National Institute of Science and Technology, Cranfield University, K.U. Leuven, Tsinghua University, University of Insbruck, University of Tartu).
Siden etableringen i 2009 har der været mange aktiviteter relateret til de ovennævnte formål. For eksempel, flere workshops og special spor (co-organiseret af Task Force'n, for eksempel workshops på Business Process Intelligence (BPI'09), BPI'10, BPI'11) og special spor på IEEEs hovedkonferencer (for eksempel CIDM'11). Viden blev formidlet via tutorials (for eksempel WCCI'10 og PMPM'09), sommerskoler (ESSCaSS'09, ACPN'10, CICH'10 osv.) og videoer (jf. www.processmining.org) og flere publikationer inklusiv den første bog om process mining, der er udgivet for nyligt af Springer. The Task Force on Process Mining har også sam-organiseret det første Business Process Intelligence Challenge (BPIC'11): en konkurrence hvor deltagerne fik til opgave at trække meningsfuld viden ud af en stor og kompleks hændelses log. I 2010 fik the Task Force on Process Mining standardiseret XES (www.xes-standard.org), et standard log format som er extensible og understøttet af OpenXES library (www.openxes.org), og at værktøjer såsom ProM, XESame, Nitro, etc.

Som læser inviteres du til at besøge www.win.tue.nl/ieetfpm/ for mere information om the Task Force's aktiviteter.

2. Process Mining: State of the Art

Process Mining bog

www.processmining.org/book/
W.M.P. van der Aalst. Process Mining: Discovery, Conformance and Enhancement of Business Processes. Springer-Verlag, Berlin, 2011.

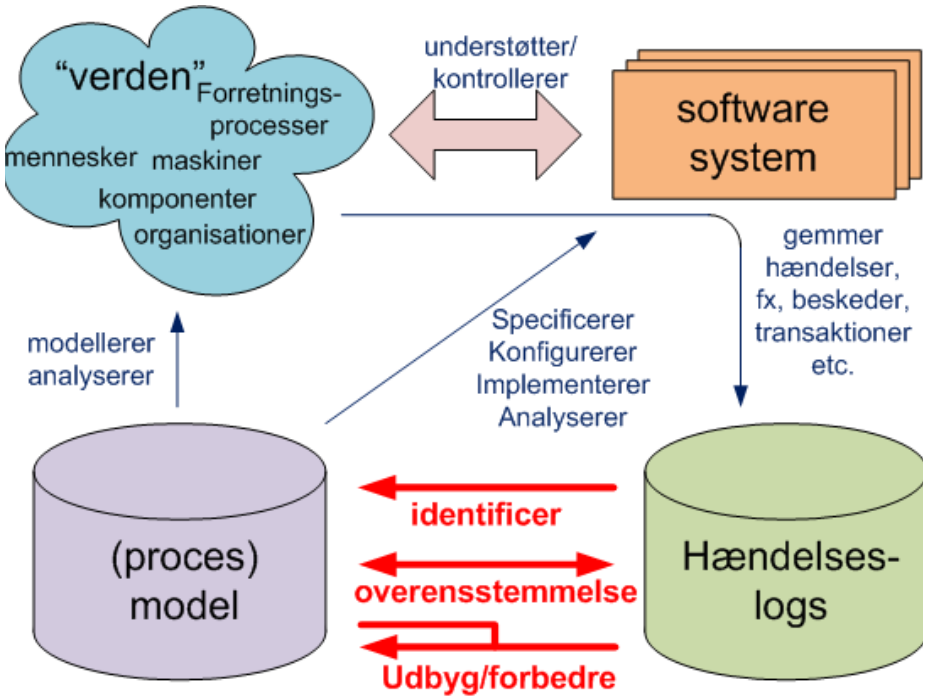


De ekspanderende kapaciteter informationssystemer og andre systemer, der er afhængige af beregninger, er godt karakteriseret af Moores lov. Gordon Moore, medstifter af Intel, forudsagde i 1965, at antallet af komponenter i integrerede kredsløb ville fordobles hvert år. I løbet af de sidste halvtreds år har væksten virkelig været eksponentiel, om end i et lidt langsommere tempo.

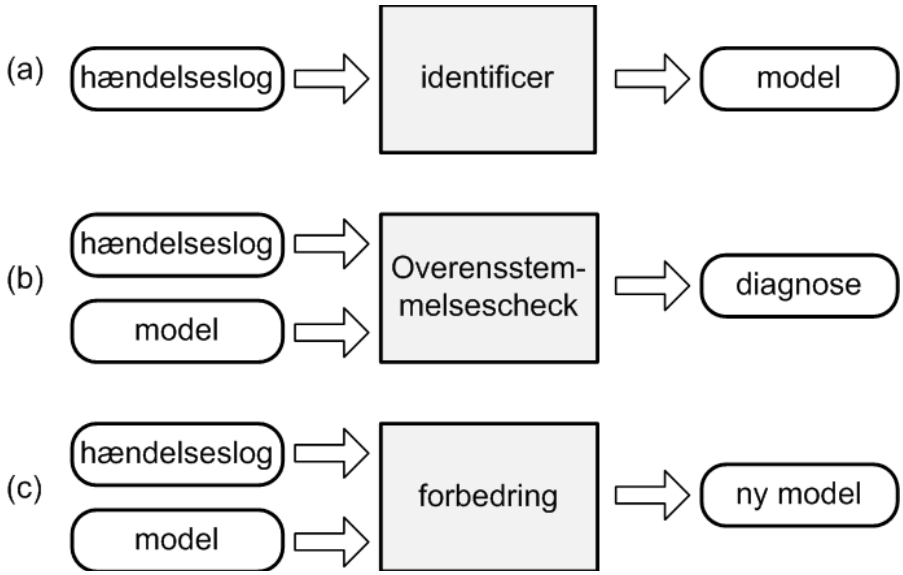
Disse fremskridt har resulteret i en spektakulær vækst i det "digitale univers" (dvs. alle data opbevares og / eller udveksles elektronisk). Desuden, bliver det digitale og det virkelige univers fortsat mere og mere på linje med hinanden.

Væksten i et digitalt univers, der er godt afstemt med processer i organisationer, gør det muligt at gemme og analysere hændelser. Hændelser kan variere fra hævning af kontanter fra en pengeautomat, en læge der justerer en røntgen maskine, en borger der ansøger om et kørekort, indsendelse af en selvangivelse og en rejsendes modtagelse af en e-billet. Udfordringen er at udnytte data hændelser på en meningsfuld måde, for eksempel, at give indsigt, identificere flaskehalse, forudse problemer, overtrædelser, anbefale modforanstaltninger, og strømline processer. Proces mining forsøger netop at muliggøre dette.

Udgangspunktet for proces mining er en hændelseslog. Alle proces mining teknikker antager, at det er muligt at gemme sekventielle hændelser, sådan at hver hændelse refererer til en aktivitet (dvs. et veldefineret trin i en proces) og er relateret til en bestemt case (dvs. en proces instans). Hændelseslogfilerne kan lagre yderligere oplysninger om hændelser. I virkeligheden kan process mining teknikker, når det er muligt, bruge ekstra oplysninger såsom ressourcen (dvs. person eller enhed) der har udført eller initieret aktiviteten, hændelsens tidsstempel, eller dataelementer gemt med hændelsen (fx, størrelsen på en ordre).
Som vist i fig. 2, kan hændelseslogfiler bruges til at foretage tre typer af proces mining. Den første type proces mining er eksplorativ. En eksplorativ teknik tager en hændelseslog, og identificerer en model uden brug af nogen form for a-priori information. Eksplorativ proces mining er den mest fremtrædende proces mining teknik. For mange organisationer er det overraskende at se, at de eksisterende teknikker faktisk er i stand til at finde rigtige processer blot baseret på eksempelvis hændelser i hændelseslogs. Den anden type proces mining er overensstemmelse check. Her sammenlignes en eksisterende proces-model med en hændelseslog af den samme proces. Overensstemmelse check kan bruges til at kontrollere, om virkeligheden, som registreres i loggen, svarer til modellen og vice versa.



Figur 2: Positionering de tre basale process mining typer: (a) *identificering*, (b) *overensstemmelse check* og (c) *forbedring*.



Figur 3: De tre basale process mining typer forklaret som input og output: (a) *identificering*, (b) *overensstemmelse check* og (c) *forbedring*.

Bemærk, at forskellige typer af modeller kan overvejes: overensstemmelse check kan anvendes til procesorienterede modeller, organisatoriske modeller, deklarative procesmodeller, forretningsregler / politikker, love etc.

Den tredje type proces mining er forbedringer. Her er det tanken at udvide eller forbedre en eksisterende proces model ved brug af oplysninger om den faktiske proces, registreret i nogle hændelseslog. Eftersom overensstemmelses check måler overensstemmelsen mellem model og virkelighed, så sigter denne tredje type proces mining på at ændre eller udvide a-priori modellen. For eksempel ved at bruge tidsstempler i hændelses-loggen, kan man udvide modellen til at få vist flaskehalse, serviceniveau, gennemløbstider og frekvenser.

Figur 3 beskriver de 3 typer af process mining i form af input og output. Teknikker til identificering af proces model tager en hændelseslog, og producerer en model. Den identificerede model er typisk en proces model (for eksempel et Petri net, BPMN, EPC eller UML aktivitets diagram). Men modellen kan også beskrive andre perspektiver (fx et socialt netværk).

Overensstemmelse check teknikker kræver en hændelseslog og en model som input. Outputtet består af diagnostiske oplysninger, der viser forskelle og fællestræk mellem model og hændelseslog. Teknikker til model forbedring (reparation eller udvidelse) har også brug for en hændelseslog og en model som input. Outputtet er en forbedret eller udvidet model. Process mining kan dække forskellige perspektiver.

Kontrol-flow perspektivet fokuserer på kontrol-flow, dvs. rækkefølgen af aktiviteter. Målet med process mining af dette perspektiv er at finde en god karakteristik af alle de mulige proces veje. Resultatet udtrykkes typisk i form af et Petri net eller en anden proces notation (for eksempel EPC, BPMN eller UML aktivitets diagrammer).

Det organisatoriske perspektiv fokuserer på information om ressourcer gemt i loggen, dvs, hvilke aktører (fx, mennesker, systemer, roller eller afdelinger) der er involveret, og hvordan er de relateret. Målet er enten at strukturere organisationen ved at klassificere mennesker i form af roller og organisatoriske enheder eller at vise det sociale netværk.

Sags (case) perspektivet fokuserer på egenskaberne ved sager (cases). Naturligvis kan en sag karakteriseres ved dens vej i processen eller af de aktører, der arbejder med den Den kan dog også være kendetegnet ved værdierne af de tilhørende dataelementer. For eksempel, hvis en sag repræsenterer en ordre genopfyldning, kan det være interessant at kende leverandøren eller antallet af bestilte varer. Tidsperspektivet omhandler timingen og hyppigheden af hændelser. Når hændelser bærer tidsstempler, er det muligt at opdage flaskehalse,

Process Mining karakteristik:

1. Process mining er ikke begrænset til identificering af kontrol-flow

Identificeringen af proces modeller ud fra hændelseslog giver grobund for andre ideer fra både udøvere og akademikere. Derfor, bliver kontrol-flow identificering ofte set som som den mest spændende del af proces mining. Process mining er dog ikke begrænset til identificeringen af kontrol-flow. På den ene side er identificering blot en af de tre basale former (identificering (eksplorativ), overensstemmelses check og forbedring). På den anden side er scopet ikke begrænset til kontrol-flow - organisations-, sags -(case) og tidsperspektivet spiller også en stor rolle.

2. Process mining er ikke blot en anden form for data mining.

Process mining kan betragtes som det "missing link" mellem data mining og traditionel model drevet BPM. De fleste data mining teknikker er ikke procesorienterede overhovedet. Proces modeller der potentielt viser parallelle sammenhænge er ikke kompatible med simple data mining strukturer såsom beslutningstræer og associerede regler. Derfor er der brug for helt nye typer repræsentationer og algoritmer.

3. Process mining er ikke kun til offline analyser.

Process mining teknikker udtrækker viden fra historiske data. Selvom "post mortem" data benyttes, kan resultaterne tillægges aktuelle cases. For eksempel kan færdiggørelsestiden, for en delvist håndteret kunde ordre, forudses ved brug af en identificeret proces model.

måle serviceniveauet, monitorere udnyttelsen af ressourcerne og forudsige den resterende proces tid på igangværende sager (cases).
Der er nogle typiske misforståelser i relation til process mining. Nogle leverandører, analytikere og forskere begrænser scopet for process mining til at være en særlig data mining teknik til identificering af processen, der kun kan bruges til offline analyser. Dette er ikke tilfældet, derfor lægger vi vægt på de tre karakteristika i boksen på forrige side.
For at positionere process mining bruger vi Business Process Management (BPM) livscyklussen som er vist i fig. 4. BPM livscyklussen viser en forretningsproces syv faser og dens tilsvarende informationssystem (er). I (re-) design fase laves en ny proces model, eller en eksisterende proces model tilpasses. I analysefasen analyseres kandidat modellen og dens alternativer. Efter (re) design fasen implementeres modellen (implementeringsfasen) eller et eksisterende system (re-) konfigureres (re-konfigureringsfasen).
I eksekveringsfasen godkendes den designede model. Under eksekveringsfasen monitoreres processen. Endvidere kan det ske, at mindre justeringer foretages uden hele processen redesignes (justeringsfasen).
I diagnose fasen bliver den godkendte proces analyseret og outputtet fra denne fase kan udløse en ny proces redesign fase.

Process mining er et værdifuldt værktøj i de fleste af faserne vist i Figur 4. Det er indlysende, at diagnose fasen kan få gavn af process mining. Men process mining er ikke begrænset til diagnose fasen. For eksempel kan process mining benyttes i eksekveringsfasen til operationel støtte. Forudsigelser og anbefalinger baseret på modeller identificeret ved historiske data kan bruges til at påvirke igangværende sager (cases). Samme typer af beslutnings støtte kan bruges til at tilpasse processer og give vejledning til (re-)konfigurering.
Hvor Figur 4. viser den overordnede BPM model og artefakter, viser Figur 5. De mulige stadier i et process mining projekt. Alle process mining projekter starter med planlægning og begrundelse for denne planlægning (stadie 0). Efter initiering af projektet er det nødvendigt at udtrække hændelsesdata, modeller, målepunkter og spørgsmål fra systemer, domæne eksperter og ledelsen (stadie 1).
Dette kræver forståelse for de data, der er til rådighed ("Hvilke data kan bruges til analysen?") og en forståelse for domænet (forretningsviden) ("Hvad er de vigtige spørgsmål?") og det resulterer i artefakterne vist i Figur 5. (for eksempel historiske data, håndlavede modeller, målepunkter og spørgsmål). I stadie 2 laves kontrol-flow modellen og linkes til hændelsesloggen. Here kan Automated Process Discovery teknikker benyttes. Den identificerede proces model kan muligvis allerede give svar på nogle af spørgsmålene og udløse redesign eller tilpasninger. Desuden kan loggen filtreres eller tilpasses ved hjælp af modellen (for eksempel fjerne sjældne aktiviteter eller afvigende sager (cases) og indsætte manglende hændelser).
Nogengange er der behov for en betydelig indsats for at korrelere hændelser, der tilhører den samme proces instans. De resterende hændelser er relateret til enheder i processen modellen. Når processen er relativt struktureret, kan kontrol-flow modellen udvides med andre perspektiver for eksempel data, tid og ressourcer i stadie 3.
Relationen mellem hændelseslog og model etableret i stadie 2 bruges til at udvide modellen (for eksempel bruges tidsstempler fra relaterede hændelser til at estimere aktiviteters ventetid). Dette kan bruges til at besvare yderligere spørgsmål, og kan udløse yderligere tiltag. Ultimativt kan modellen, konstrueret i stadie 3, bruges til operationel support (Stadie 4) viden om igangværende sager (cases). Dette kan bruges til at gribe ind, forudsige og anbefale tiltag. Stadierne 3 og 4 kan kun nås, hvis processen er tilstrækkeligt stabil og struktureret.
I øjeblikket er der teknikker og værktøjer, der kan understøtte alle stadier vist i Figur 5. Dog er process mining et relativt nyt paradigme og de fleste af de nuværende værktøjer er stadig forholdsvist umodne. Derudover, er mulige brugere ofte ikke klar over potentialet og begrænsningerne i process mining. Det er grunden til dette manifest, der giver nogle vejledende principper (jf næste sektion) og redegør for nogle udfordringer (jf side 10) for process mining brugere såvel som forskere og udviklere, som er interesserede i at fremme state-of-the-art.

Vejledende principper:

VP1: Hændelsesdata bør ses som 1. klasses borgere

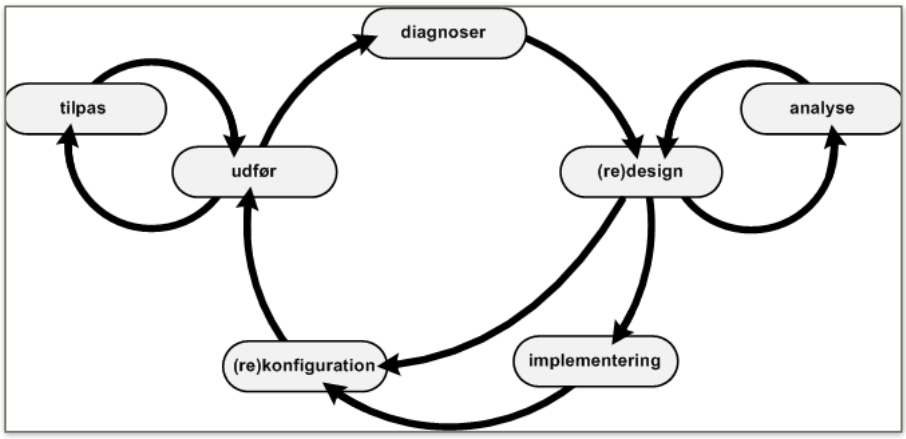
VP2: Log udtræk bør drives af spørgsmål

VP3: Samtidighed, valg og basale kontrol-flowkonstruktioner bør understøttes

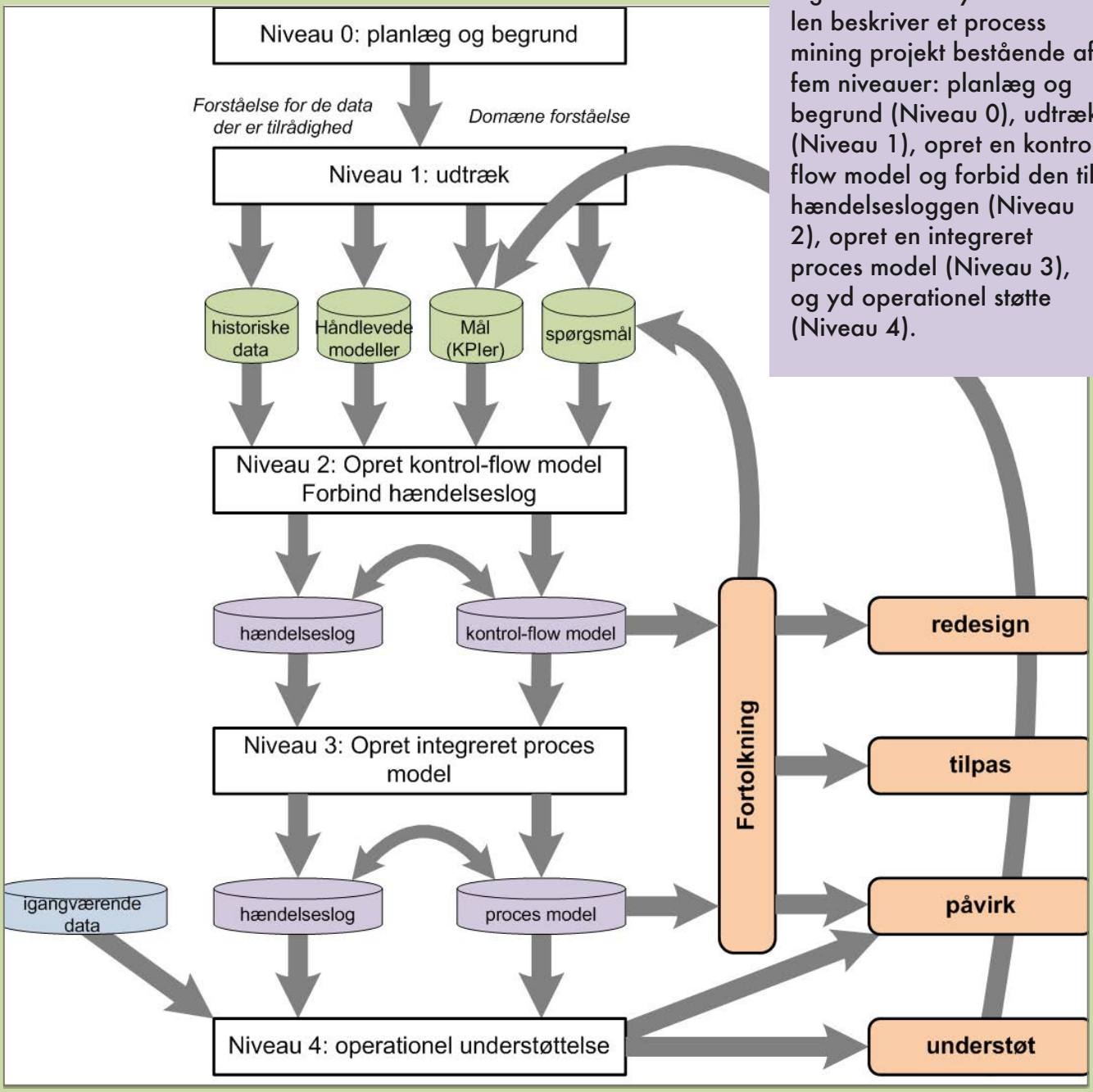
VP4: Hændelser bør relateres til model elementer

VP5: Modeller bør behandles som målrettede abstraktioner af virkeligheden

VP6: Process mining bør være en kontinuerlig proces



Figur 4: BPM life-cycle identificerer de forskellige faser i en forretningsproces og tilhørende information system(er); process mining (potentielt) spiller en rolle i alle faser (bortset fra implementeringsfasen).



Figur 5: L* life-cycle modellen beskriver et process mining projekt bestående af fem niveauer: planlæg og begrund (Niveau 0), udtræk (Niveau 1), opret en kontrol-flow model og forbid den til hændelsesloggen (Niveau 2), opret en integreret proces model (Niveau 3), og yd operationel støtte (Niveau 4).

3. Vejledende principper

Som med al anden ny teknologi er der nogle åbenlyse fejltagelser, der kan laves, når man benytter process mining i et virkelighedstro setup. Derfor nævner vi 6 vejledende principper for at undgå at brugere / analytikere begår disse fejltagelser.

VP1: Hændelsesdata bør ses som første klasses borgere

Udgangspunktet for al process mining aktivitet er de lagrede hændelser. Vi refererer til samlinger af hændelser som hændelseslogs, dette betyder imidlertid ikke, at hændelser behøves at være lagret i dedikerede logfiler. Hændelser kan lagres i database tabeller, message logs, mail arkiver, transaktionslogs og andre data kilder. Vigtigere end lagringsformatet er kvaliteten af disse hændelseslogs. Kvaliteten af et process mining resultat afhænger kraftigt af inputtet. Derfor bør hændelseslogs behandles som første klasses borgere i de informationssystemer, der understøtter processerne, der skal analyseres.
Uheldigvis er hændelseslogs ofte mere et "bi-produkt" brugt til debugging eller profilering. For eksempel gemmer medicinsk udstyr fra Philips Healthcare hændelser, simpelthen fordi udviklere har indsat "print udsagn" i koden. Selvom der er nogle uformelle retningslinjer for tilføjelse af sådanne udsagn i koden, er der brug for en mere systematisk tilgang for at forbedre kvaliteten af hændelseslogs. Hændelsesdata bør betragtes som første klasses borgere (i højere grad end andenrangs borgere).

Niveau	Karakteristik	Eksempel
★ ★ ★ ★ ★	Højeste niveau: Hændelsesloggens kvalitet er excellent (dvs. troværdig og komplet) og hændelser er veldefinerede. Hændelser gemmes på en automatiseret, systematisk, pålidelig og sikker måde. Privatlivs- og sikkerhedsspørgsmål er tilstrækkeligt adresseret. Desuden er hændelser gemt, og (attributterne) har tydelig semantik. Dette indebærer eksistensen af en eller flere ontologier. Hændelser og deres attributter peger på den ontologi.	Semantisk noterede logs fra BPM systemer.
★ ★ ★ ★	Hændelser gemmes automatisk og på en systematisk og pålidelig måde, dvs logs er troværdige og komplette. Ulig systemerne på niveau ★★★ , er notationer såsom proces instans explicit understøttet.	Hændelseslogs fra traditionelle BPM/workflow systemer.
★ ★ ★	Hændelser gemmes automatisk men uden en systematisk tilgang. Dog er der, til forskel for logs på niveau ★★ , en hvis garanti for at lagrede hændelser stemmer overens med virkeligheden (dvs. hændelsesloggen er troværdig men ikke nødvendigvis komplet). Tag for eksempel hændelser gemt i ERP systemer. Selvom hændelser er nødt til blive udtrukket fra mange tabeller, kan information antages at være korrekt (fx, er det sikkert at antage, at en lagret betaling i ERP systemet faktisk eksisterer og vice versa).	Tabeller i ERP systemer, hændelseslogs fra CRM systemer, transaktionslogs fra messaging systemer, hændelseslogs fra high-tech systemer, etc.
★ ★	Hændelser gemmes automatisk, dvs. som et bi-produkt fra nogle informations systemer. Dækningen varierer, dvs. ingen systematisk tilgang følges for at bestemme, hvilke hændelser der lagres. Desuden er det muligt at omgå informations systemet. Hvorved hændelser kan mangle eller ikke være gemt ordenligt.	Hændelseslogs fra dokument og product management systemer, fejllogs fra embedded systems, regneark fra serviceteknikkere, etc.
★	Laveste niveau: hændelseslogs er af dårlig kvalitet. Lagrede hændelser svarer måske ikke til virkeligheden, eller kan mangle. Hændelseslogs der laves manuelt, har typisk sådanne karakteristika.	Spor efterladt i papirdokumenter ført gennem organisationen ("gule sedler") papirbaserede journaler, etc

Tabel 1: Modenhedsniveauer for hændelseslogs.

Der findes flere kriterier til bedømmelse af kvaliteten af hændelsesdata. Hændelser bør være troværdige, for eksempel bør det være sikkert at antage, at de lagrede hændelser faktisk skete, og at hændelsernes attributter er korrekte. Hændelseslogs bør bære komplette, for eksempel, i et givent omfang, må der ikke mangle hændelser. Alle lagrede hændelser bør have veldefineret semantik. Desuden bør data være sikker i den forstand at privatlivs- og sikkerhedsmæssige spørgsmål er håndteret før hændelserne lagres. For eksempel skal aktører være klar over, hvilken slags hændelser der lagres, og hvordan de bliver brugt. Tabel 1 definerer 5 hændelseslog modenhedsniveauer rangerende fra excellent kvalitet (*****) til dårlig kvalitet (*). For eksempel ligger hændelseslogfilerne fra Philips Healthcare på niveau ***. Det vil sige at hændelserne lagres automatisk, og den registrerede adfærd svarer til virkeligheden., men der er ikke en systematisk tilgang til at tildele semantik til hændelser og sikre dækning på et bestemt niveau. Process mining teknikker kan foretages på logfiler på niveau *****, **** og ***. I princippet er det også muligt at anvende process mining på hændelseslogfiler på niveau ** og *. Dog er analysen på sådanne logfiler typisk problematiske og resultatet er ikke troværdigt. Faktisk giver det ikke mening, at anvende process mining på logfiler på niveau *.
For at få noget ud af process mining bør organisationer prøve at få hændelseslogfiler på så højt et niveau som muligt.

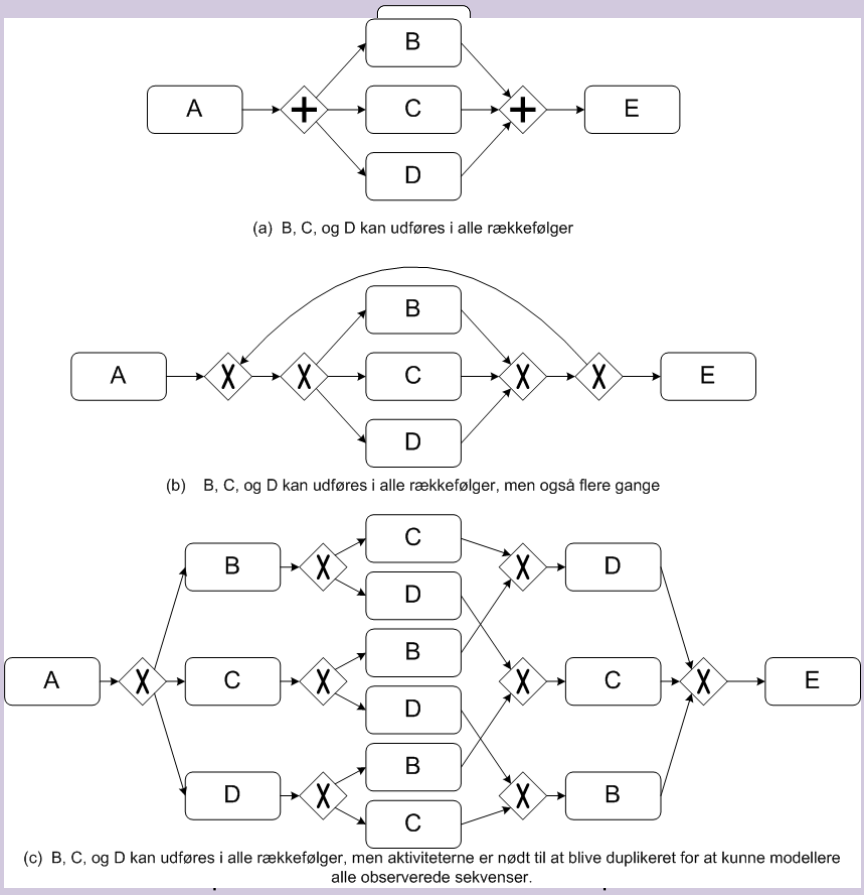
VP2: Log udtræk bør drives af spørgsmål

Som vist i Figur 5 er process mining aktiviteterne nødt til være drevet at spørgsmål. Uden konkrete spørgsmål er det meget vanskeligt at udtrække meningsfulde hændelsesdata. Forestil dig for eksempel de tusindvis af tabeller der findes i et ERP system som SAP. Uden konkrete spørgsmål er det umuligt at vælge de tabeller, der er relevante at trække data ud af. En proces model som den vist i Figur 1 beskriver livscyklussen for sager (cases) (det vil sige proces instancer) af en bestemt type. Derfor er man nødt til vælge hvilken type af sager (cases), der skal analyseres, før der bruges process mining teknikker på data. Dette valg bør være drevet af de spørgsmål, der ønskes

Besvaret og dette er måske ikke trivielt. Tænk for eksempel på håndteringen af kunde ordrer. Hver kunde ordre kan bestå af flere ordre linjer, da kunden kan bestille flere produkter i én ordre. En kunde ordre kan resultere i flere leverancer. En leverance kan henvise til ordre linjer fra flere forskellige ordrer. Der er altså en mange-til-mange relation mellem ordrer og leverancer og en en-til-mange relation mellem ordre og ordre linjer. I en database med hændelsesdata relateret til ordrer, ordre linjer og leverancer kan der identificeres forskellige proces modeller. Man kan udtrække data med det formål at beskrive livscyklussen for individuelle ordrer. Men det er også muligt at udtrække data med det formål at identificere livscyklussen for individuelle ordre linjer eller livscyklussen for individuelle leverancer.

VP3: Samtidighed, valg og andre basale kontrol-flow konstruktioner bør være understøttet

Der findes et væld af procesmodellerings sprog (for eksempel BPMN, EPC, Petri Net, BPEL og UML aktivitets diagrammer). Nogle af disse sprog indeholder mange modelleringselementer (for eksempel har BPMN mere end 50 forskellige grafiske elementer), hvorimod andre er meget basale (for eksempel har Petri Nets kun 3 forskellige elementer: steder, overgange og buer. Kontrol-flow beskrivelsen er rygraden i enhver proces model. De basale workflow konstruktioner (også kendt som *patterns*) der understøttes af alle mainstream sprog er sekvens, parallelitet (AND-splits/joins), valg (XOR-splits/joins) og loops. Disse *patterns* skal naturligvis også understøttes af process mining teknikker. Men nogle teknikker er ikke i stand til at håndtere samtidighed og understøtter kun Markovkæder / transitionssystemer. Figur 6 viser effekten af at bruge proces mining teknikker, der ikke er i stand til at identificere samtidighed (ingen AND-split/joins). Tænk på en hændelseslog L= {<A, B, C, D, E>, <A, B, D, C, E>, <A, C, B, D, E>, <A, C, D, B, E>, <A, D, B, C, E>, <A, D, C, B, E>}. L indeholder altså sager der starter med A og sluttet med E. Aktiviteterne B, C og D udføres i en hvilken som helst rækkefølge imellem A og E. BPMN



Figur 6: Eksempel der illustrerer problemer når samtidighed (dvs. AND-splits/joins) ikke kan udtrykkes direkte. I eksemplet er bare tre aktiviteter samtidige (B, C, and D). Tænk på proces modellerne når der er 10 samtidige aktiviteter (2^{10}=1024 tilstande og10! = 3,628,800 mulige udførbare sekvenser).

modellen i Figur 6 (a) viser en kompakt repræsentation af den underliggende proces ved brug af to AND gateways. Antag at process mining teknikken ikke understøtter AND gateways. I dette tilfælde er de to andre BPMN modeller i Figur mulige kandidater for en repræsentation. BPMN modellen i Figur 6(b) er kompakt, men tillader for mange forløb (fx er sager (cases) som <A, B, B, B, E> mulige ifølge modellen, men er ikke sandsynlige ifølge hændelses-loggen.). BPMN modellen i Figur 6(c) tillader alle sager (cases) i L, men definerer alle sekvenser eksplicit, så det er ikke længere en kompakt repræsentation og loggen. Eksemplet viser, at for virkelighedstro modeller, der har et utal af mulige parallelle aktiviteter, resulterer det i modeller med voldsom *underfitting* (det vil sige mulighed for for mange forløb) og/eller ekstremt kompleksitet, hvis parallelitet ikke understøttes. Som illustreret i Figur 6 er det vigtigt, i det mindste, at understøtte de basale workflow patterns. Udover de basale workflow patterns er det ønskeligt at understøtte OR-splits/joins, fordi de giver mulighed for en kompakt repræsentation af inkluder-ende beslutninger og delvise synkroniseringer.

VP4: Hændelser bør være relateret til model elementer

Som antydet i sektion 2 er det en misforståelse, at process mining er begrænset til kontrol-flow identificering. Som vist i Figur 1 kan den identificerede proces model dække over flere perspektiver (organisatorisk perspektiv, tidsperspektiv, data perspektiv og så videre). Endvidere er identificering bare en af tre typer af process mining som vist i Figur 3.

De andre typer af process mining (overensstemmelse check og forbedring) hviler tungt på relationen mellem elementer i modellen og hændelser i loggen. Denne relation kan bruges til at "afspille" hændelsesloggen på modellen. Afspilning kan bruges til at afdække forskelle mellem en hændelseslog og en model, for eksempel er nogle af hændelserne i loggen ikke mulige ifølge modellen. Teknikker til overensstemmelse check kvantificerer og diagnosticerer sådanne forskelle. Tidsstempler i hændelsesloggen kan bruges til at analysere de midlertidige forløb under afspilning. Tidsforskelle mellem løst relaterede aktiviteter kan bruges til at tilføje forventede ventetider til modellen. Disse eksempler viser, at sammenhængen mellem hændelser i loggen og elementer i modellen fungerer som et start punkt for forskellige analyse typer.
I nogle tilfælde er det ikke helt trivielt at etablere sådanne relationer. For eksempel kan en hændelse henvise til to forskellige aktiviteter, eller det er svært at se hvilken aktivitet den refererer til. Sådanne tvetydigheder skal fjernes for at kunne fortolke process mining resultaterne korrekt. Udover problemet med at relatere hændelser til aktiviteter, er der problemet med at relatere hændelser til proces instancer. Dette kaldes normalt for hændelseskorrelation.

VP5: Modeller bør behandles som målrettede abstraktioner af virkeligheden

En model udledt fra hændelsesdata giver et billede af virkeligheden. Sådan et billede bør give en målrettet abstraktion af de forløb, der er gemt i hændelsesloggen. Da det er en hændelseslog, kan der være flere "billeder" af virkeligheden, der er brugbare. Derudover kan forskellige interessenter kræve forskellige "billeder". Faktisk bør modeller identificeret på baggrund af hændelseslogfiler ses som "kort" (ligesom geografiske kort). Dette princip giver vigtige indsigter, hvoraf to er beskrevet i den resterende del.
For det første er det vigtigt at bemærke, at der ikke findes et "rigtigt kort" for et givent geografisk område. Alt afhængig af formålet, er der forskellige kort: vejkort, vandretur kort, cykel kort og så videre. Alle disse kort viser et billede af den samme virkelighed, og det ville være absurd at antage, at der findes noget der kan kaldes det "perfekte kort". Det samme gælder for proces modeller: modellen bør ligge vægt på de ting, der er relevante for en bestemt type bruger. Identificerede modeller kan fokusere på forskellige perspektiver (kontrol-flow, data flow, tid, ressourcer, omkostninger og så videre) og bør vise disse på forskellige abstraktionsniveauer og detaljegrader. For eksempel kan en leder ønske at se en overordnet uformel proces, der fokuserer på omkostningerne, hvorimod en proces analytiker måske ønsker at se en detaljeret proces model med fokus på afvigelser fra det normale flow. Bemærk også, at forskellige interessenter måske ønsker at se en proces på forskellige niveauer: strategisk niveau (beslutninger på dette niveau har langsigtede virkninger, og er baseret på aggregerede hændelsesdata over en længere periode), taktiske niveau (beslutninger på dette niveau har mellemlangtidsvirkninger, og er for det meste baseret på de seneste data) og operationelt niveau (beslutninger på dette niveau har umiddelbare virkninger og er baseret på hændelsesdata med relation til igangværende sager (cases)).
For det andet er det nyttigt at indføre ideer fra kartografi, når det kommer til at producere forståelige kort. For eksempel abstraherer vejkort fra mindre vigtige veje og byer. Mindre vigtige ting er enten udeladt eller dynamisk grupperet i aggregerede former (for eksempel er gader og forstæder sammensmeltet til byer). Kartografer fjerner ikke alene irrelevante detaljer, men bruger også farver til at fremhæve vigtige funktioner. Derudover har grafiske elementer en bestemt størrelse for at angive deres betydning (fx kan størrelserne af linjer og prikker variere). Geografiske kort har også en klar fortolkning af x-aksen og y-aksen, dvs. at udformningen af et kort ikke er vilkårlig, da koordinaterne for elementer har en betydning. Alt dette står i skarp kontrast til mainstream procesmodeller, der typisk ikke anvender farve, størrelse og placering af funktioner, der gør modellerne mere forståelige. Dog kan ideer fra kartografi nemt indarbejdes i konstruktionen af identificerede proces kort. For eksempel kan størrelsen af en aktivitet bruges til at reflektere dens frekvens eller andre faktorer, der angiver dens betydning (for eksempel omkostninger eller ressourceforbrug). Bredden af en bue kan afspejle betydningen af den tilsvarende kausale afhængighed, og farvning af buer kan bruges til at fremhæve flaskehalse.
Ovennævnte observationer viser, at det er vigtigt at vælge den rigtige repræsentation og fin-tune den til modtagerne. Dette er vigtigt for at kunne visualisere resultaterne til slutbrugerne og guide identificeringsalgoritmerne hen imod passende modeller (se også udfordring U5).

VP6: Process mining bør være en kontinuerlig proces

Process mining kan hjælpe med at give meningsfulde "kort", som er direkte forbundet til hændelsesdata. Både historiske hændelsesdata og nuværende data kan projiceres ind i sådanne modeller. Derudover ændrer processerne sig, alt imens de analyseres. Givet processers dynamiske natur anbefales det ikke at se process mining som en engangs aktivitet. Målet bør ikke være at lave en færdig statisk model, men at puste liv ind i proces modeller således at brugere og analytikere opmuntres til at se på dem dagligt.
Sammenlign dette med grafiske overblik frembragt ved brug af geo-tagging. Der findes tusindvis af grafiske overblik lavet af Google Maps (for eksempel applikationer, der lægger information om trafik forhold, fast ejendom, fastfood restauranter eller biograf forestillinger ind på et valgt kort. Folk kan problemfrit zoome ind og ud ved hjælp af sådanne kort og interagere med dem (fx er trafikpropper projiceret ind på kort og brugeren kan vælge at se detaljer for et særligt problem). Det bør også være muligt at gennemføre process mining baseret på real-time hændelsesdata. Ved brug af "kort" metaforen kan vi forestille os at GPS koordinater kan projiceres in på kort i realtid. Analogt til bilers navigationssystemer, kan process mining hjælpe slutbrugere (a) ved at navigere gennem processer (b) ved at projicere dynamisk information ind på proces kort (for eksempel vise "trafik propper" i forretningsprocesser) og (c) ved at give bud på forudsigelser vedrørende igangværende sager (cases) (for eksempel estimere "ankomsttiden" for en forsinket sag(case)). Disse eksempler demonstrerer, at det er ærgerligt, hvis ikke man bruger proces modeller mere aktivt. Derfor bør process mining betragtes som en iterativ proces, der giver information om relevante tiltag på varierende tidshorisonter (minutter, timer, dage, uger og måneder).

4. Udfordringer

Process mining er et vigtigt værktøj for moderne organisationer, der er nødt til at håndtere ikke-trivielle operationelle processer. På den ene side er der en utrolig vækst i hændelsesdata. På den anden side er processer og information nødt til at blive ensrettet for at imødese kravene til overensstemmelse (check), effektivitet og kundeservice. Trods anvendeligheden af process mining er der stadig store udfordringer, der skal løses; disse illustrerer, at process mining er en ny disciplin på vej frem. I den resterende del lister vi nogle af disse udfordringer op. Det er ikke tænkt som en komplet liste og, over tid, vil nye udfordringer dukke op eller eksisterende vil forsvinde i kraft af, at process mining modnes.

U1: Find, sammenlæg og rens hændelsesdata

Det kræver stadig en betydelig indsats at udtrække hændelsesdata, der er egnet til process mining. Typisk er der flere hurdler, der skal overkommes:

- Data kan være distribueret ud over flere kilder. Denne information er nødt til at blive sammenlagt. Der er en tendens til, at det er problematisk når forskellige identifikatorer anvendes i de forskellige datakilder. For eksempel bruger et system navn og fødselsdato til at identificere en person, mens et andet system bruger personens CPR-nummer.
- Hændelsesdata er oftere "objekt orienteret", end det er proces orienteret. For eksempel individuelle produkter, paller og containere med RFID-tags og lagrede hændelser henviser til disse tags. Men for at overvåge en bestemt kunde ordre skal sådanne objekt-orienterede hændelser sammenlægges og forbehandles.
- Hændelsesdata er måske ikke komplette. Et almindeligt problem er, at hændelser ikke peger eksplicit på en proces instans. Ofte er det muligt at udlede den information, men det kan kræve en betragtelig indsats. Tidsdimensionen kan også mangle for nogle af hændelserne. Man er måske nødt til at tilpasse tidsstempler for være i stand til at bruge den tidsinformation, der er til rådighed.
- En hændelseslog kan indeholde "outliers", dvs. usædvanlig adfærd også omtalt som støj. Hvordan defineres "outliers"? Hvordan identificeres sådanne "outliers"? Disse spørgsmål skal besvares for at kunne rense hændelsesdata.
- Logs kan indeholde hændelser på forskellige granuleringsniveauer. I hændelsesloggen fra et hospitals informations system kan hændelser henvise til simple blodprøver eller til komplekse kirurg procedurer. Tidsstempler kan også have forskellige granuleringsniveauer rangerende fra millisekunders præcision(28-9-2011:h11m28s32ms342) til overordnet dato information (28-9-2011).
- Hændelser sker i en bestemt kontekst (vejr, arbejdsbelastning, ugedag og så videre). Denne kontekst kan forklare bestemte fænomener, for eksempel at svartiden er længere end sædvanligt grundet igangværende arbejde eller ferier. For analyser er det ønskeligt at indarbejde denne kontekst. Dette indebærer sammenlægning af hændelsesdata med kontekstuelle data. Her er det "forbandelsens dimensionalitet" dukker op efterhånden som analysen bliver umedgørlig, i takt med at du tilføjer for mange variabler.

Bedre værktøjer og metoder er nødvendige for at kunne adressere ovennævnte problemer. Derudover, som tidligere antydet, er organisationer nødt til at behandle hændelseslogs som første klasses borgere i højere grad end som et bi-produkt. Målet er at opnå ***** hændelseslogs (se Tabel 1). Her er erfaringerne i forbindelse med datawarehousing nyttige til at sikre høj kvalitet i hændelseslogfiler. For eksempel kan simpel data kontrol under indtastning af data bidrage til at reducere andelen af ukorrekte hændelsesdata betydeligt.

U2: Håndtere komplekse hændelseslogs der har forskelligartede karakteristika

Hændelseslogs kan have meget forskellige karakteristika. Nogle hændelseslogs kan være så ekstremt store, at det er svært at håndtere dem, hvorimod andre hændelseslogs kan være så små, at der ikke er nok data til rådighed til at give pålidelige konklusioner. I nogle områder, registreres ufattelige mængder af hændelser. Derfor er der behov for en yderligere indsats for at forbedre performance og skalerbarhed. For eksempel overvåger ASML løbende alle sine "wafer" scannere. Disse "wafer" scannere anvendes af forskellige organisationer (for eksempel Samsung og Texas Instruments) til fremstilling af chips (ca. 70% af alle chips er fremstillet ved hjælp af ASMLs "wafer" scannere). Eksisterende værktøjer kan have svært ved at håndtere petabytes af data indsamlet i sådanne områder. Udover antallet af lagrede hændelser er der andre karakteristika såsom gennemsnitlig antal hændelser pr sag (case), lighed blandt sager (cases), antallet af unikke forløb og antallet af unikke veje. Forestil dig en hændelseslog L1 med følgende karakteristika: 1000 sager (cases), i gennemsnit 10 hændelser pr sag (case) og lav grad af variation (for eksempel følger mange sager (cases) den samme vej). Hændelseslog L2 indeholder blot 100 sager (cases) men i gennemsnit er der 100 hændelser pr sag (case) og alle sager (cases) følger en unik vej. Det er klart at L2 er meget sværere at analysere end L1 selvom de to logs har samme størrelse (ca. 10.000 hændelser). Eftersom hændelseslogfiler kun indeholder afprøvede forløb, bør de ikke betragtes som værende komplette. Proces mining teknikker er nødt til at håndtere ufuldstændighed ved hjælp af en

Udfordringer:

U1: Find, sammenlæg og rens hændelsesdata

U2: Håndtere komplekse hændelseslogs der har forskelligartede karakteristika

U3: Opret repræsentative benchmarks

U4: Håndter koncept forskydning

U5: Forbedre repræsentativ bias brugt til proces identificering

U6: Balancere mellem kvalitetskriterier såsom fitness, Enkelhed, præcision og generalisering

U7: Tværorganisatorisk mining

U8: Yde operationel støtte

U9: Kombinere process mining med andre analysetyper

U10: Forbedre Usability for ikke-Eksperter

U11: Forbedre forståelighed for ikke-eksperter

åben verden antagelse": den kendsgerning, at noget ikke skete betyder ikke, at det ikke kan ske. Dette gør det udfordrende at håndtere små hændelseslogs med stor variation.
Som før nævnt indeholder nogle hændelseslogs hændelser på et meget lavt abstraktionsniveau. Disse logs har en tendens til at være ekstremt store og de individuelle lav-niveau hændelser er ikke særligt interessante for interessenterne. Derfor kan man aggregere lav-niveau hændelser ind i overordnede hændelser. For eksempel når man analyserer diagnostik- og behandlingsprocesserne for en bestemt gruppe patienter, er man måske ikke interesseret i de individuelle tests lagret i hospital laboratoriets IT systemer.
På dette tidspunkt er organisationer nødt til at benytte en trial-and-error tilgang, for at se om en hændelseslog er velegnet til process mining. Derfor bør værktøjer give mulighed for en hurtig feasibility test for et givet data sæt. En sådan test bør indikere potentielle performance problemer og advare, hvis logs er for langt fra at være komplette, eller er for detaljerede.

U3: Opret repræsentative benchmarks

Process mining er en teknologi på vej fremad. Det forklarer hvorfor der stadig mangler gode benchmarks. For eksempel findes der utallige proces identificerings teknikker, og forskellige leverandører tilbyder forskellige produkter, men der er ingen konsensus omkring kvaliteten af disse teknikker. Selvom der er stor forskel på funktionalitet og performance, er det svært at sammenligne de forskellige teknikker og værktøjer. Derfor er det nødvendigt, at der udvikles gode benchmark med eksempel datasæt og med repræsentative kvalitetskriterier. For klassisk data mining teknikker findes der mange gode benchmarks. Disse benchmarks har stimuleret værktøjs udbydere og forskere til at forbedre deres teknikkers performance. I process minings tilfælde er det mere udfordrende. For eksempel er den relationelle model, der blev introduceret af Codd i 1969 simpel og vidt understøttet. Som resultat af det, kræver det ikke særligt meget at konvertere data fra en database til en anden, og der er ingen forståelses problemer. For processer mangler sådan en simpel model. Standarder for proces modellering er meget mere komplicerede og med få leverandør, der understøtter det samme sæt af koncepter. Processer er simpelthen mere komplekse and tabulære data. Ikke desto mindre er det vigtigt at lave repræsentative benchmarks. Noget initialt arbejde er allerede til rådighed. For eksempel er der forskellige målinger for måling af kvaliteten af process mining resultaterne (fitness, enkelhed, præcision og generalisering). Derudover er der flere hændelseslogs, der er offentligt tilgængelige (jf. www.processmining.org). Se for eksempel hændelsesloggen der blev brugt i forbindelse med den første Business Process Intelligence Challenge (BPIC'11) organiseret af Task Force'n (jf. Doi:10.4121 / UUID: d9769f3d-0ab0-4fb8-803b-0d1120 ffcf54).
På den ene side burde der være benchmarks baseret på virkelighedstro datasæt. På den anden side er der brug for at oprette syntetiske datasæt, der fanger bestemte karakteristika. Sådanne syntetiske datasæt bidrager til at udvikle process mining teknikker, der er skræddersyet til ufuldstændige hændelseslogfiler, støjende hændelseslogs, eller specifikke populationer af processer.
Udover oprettelsen af repræsentative benchmarks er der også behov for mere konsensus omkring kriterierne, der benyttes til at bedømme kvaliteten af et process mining resultat (se også udfordring U6). Endvidere kan krydsvaliderings teknikkerne fra data mining adopteres til at bedømme resultatet. Tænk for eksempel på k-fold check. Man kan opsplitte hændelsesloggen i k dele. K-1 dele kan bruges til at lære en proces model at kende, og overensstemmelseskontrol teknikker kan bruges til at bedømme resultatet for den resterende del. Dette kan gentages k gange, og giver nogle indsigter omkring modellens kvalitet.

U4: Håndter koncept forskydning

Begrebet *koncept forskydning* henviser til situationen, hvor processen ændrer sig, mens den bliver analyseret. For eksempel kan det være at to aktiviteter udføres parallelt i starten af hændelsesloggen, men hvor disse to aktiviteter bliver sekventielle i slutningen af hændelsesloggen. Processer kan ændre sig grundet periodiske / sæsonbetonede ændringer (for eksempel er der i december mere "efterspørgsel" eller "på fredage aftener er der færre medarbejdere til stede") eller det kan skyldes ændrede betingelser (for eksempel "markedet bliver mere konkurrencepræget"). Sådanne ændringer påvirker processerne, og det er vitalt at opdage det og analysere dem. Konceptet forskydning i en proces kan identificeres ved at opdele hændelsesloggen i mindre dele og analysere "fodsporerne" i de mindre logs. Sådan en "Anden omgang" analyse kræver meget mere hændelsesdata.
Ikke desto mindre befinder meget få processer sig i en statisk udgave og forståelse for konceptet forskydning er af afgørende betydning for håndtering af processer. Derfor er yderligere forskning og værktøjs understøttelse nødvendig for i tilstrækkelig grad at analysere konceptet forskydning.

U5: Forbedre repræsentativ bias brugt til proces identificering

En proces identificerings teknik genererer en model ved brug af et bestemt sprog (for eksempel BPMN eller Petri nets). Dog er det vigtigt at adskille visualiseringen af resultatet fra repræsentationen benyttet under selve identificerings processen. Valget at sprog indbefatter ofte flere implicitte antagelser. Det begrænser søge området; processer, der ikke kan repræsenteres af målsproget, kan ikke identificeres. Denne såkaldte "repræsentative bias" benyttet under identificerings processen bør være et bevidst valg og bør ikke (kun) være foranlediget af den foretrukne grafiske repræsentation.
Betragt for eksempel Figur 6: Om målsproget tillader parallelitet eller ej kan have betydning for både visualiseringen af den identificerede model og algoritmens klassificering af modellen. Hvis den repræsentative bias ikke tillader parallelitet (Figur 6(a) er ikke mulig) og ikke tillader at flere aktiviteter har samme label (Figur 6(c) er ikke mulig), så er kun problematiske modeller som vist i Figur 6(b) mulige. Dette eksempel viser, at et en mere forsigtig og raffineret udvælgelse af repræsentativ bias er nødvendig.

U6: Balancere mellem kvalitetskriterier såsom fitness, Enkelhed, præcision og generalisering

Hændelseslogs er ofte langt fra at være komplette, det vil sige at kun eksempler på forløb er givet. Proces modeller tillader typisk et exponentielt eller uendeligt antal forløb af forskellig spor (i tilfælde af loops). Derudover kan nogle spor have meget mindre sandsynlighed end andre. Derfor er det urealistisk at antage at alle mulige spor er repræsenteret i hændelsesloggen. For at illustrere at det er upraktisk at tage komplette logs for givet, så tænk på en proces der består af omkring 10.000 sager (cases). Det totale antal af mulige kombinationer i en model med 10 parallelle aktiviteter er 10! = 3.628.800. Det er umuligt at have alle kombinationer repræsenteret i loggen, da der er meget færre sager (cases) (10.000) end de potentielle spor (3.628.800). Selv hvis der er millioner af sager (cases) i loggen er det ekstremt usandsynligt, at alle mulige varianter er repræsenteret. En yderligere komplikation er at nogle alternativer er mindre hyppige end andre. Disse kan betragtes som "støj". Det er umuligt at bygge en fornuftig model for sådanne "støjende" forløb. Den identificerede model er nødt til at abstrahere fra dette; det er bedre at undersøge lavt hyppige forløb ved brug af overensstemmelse check. Støj og ufuldstændighed gør proces identifikation til et vanskeligt problem. Faktisk er der fire konkurrerende kvalitets dimensioner: (a) Fitness, (b) Enkelthed, (c) præcision og (d) generalisering. En model med god *Fitness* understøtter de fleste af forløbene i hændelsesloggen. En model har en perfekt Fitness, hvis alle spor i loggen kan gentages af modellen fra start til slut. Den simpleste model der kan forklare adfærden, der kan ses i loggen er den bedste model. Dette princip er kendt som Occam's Razor. Fitness og enkelthed alene er ikke nok til at bedømme en identificeret models kvalitet. Det er for eksempel meget nemt at konstruere et ekstremt simpel Petri Net ("flower model"), der er i stand til at gentage alle spor i hændelsesloggen (men også en hvilken som helst anden hændelseslog, der refererer til samme sæt af aktiviteter). Ligeledes, er det ikke ønskeligt at have en model, der kun tillader den eksakte adfærd som ses i hændelsesloggen. Husk at loggen kun indeholder eksempler på adfærd og at mange mulige spor måske ikke er set endnu. En model er præcis, hvis den ikke tillader "for meget" adfærd. Det er tydeligt at "flower modellen" mangler præcision. En model der ikke er præcis er "underfitting". "Underfitting" er problemet, hvor en model over-generaliserer eksempel adfærden i loggen (dvs. modellen tillader adfærd som er meget anderledes end det, der var set i loggen). En model bør generalisere og ikke begrænse adfærden til de eksempler, der er set i loggen. En model der ikke generaliserer er "Overfitting". Overfitting er det problem, at en meget specifik model genereres, hvor det er tydeligt at loggen kun indeholder eksempler på adfærd (dvs., modellen forklarer den pågældende log, men en anden log for samme proces vil måske generere en helt anden proces model). At balancere Fitness, enkelthed, præcision og generalisering er udfordrende. Det er grunden til, at de fleste kraftfulde proces identificerings teknikker giver forskellige parametre. Der er nødt til at blive udviklet bedre algoritmer for bedre at balancere de fire konkurrerende kvalitets dimensioner. Desuden bør alle benyttede parametre være forståelige for brugerne.

U7: Tværorganisatorisk mining

Traditionelt set er process mining blevet benyttet indenfor en enkelt organisation. Men efterhånden som service teknologi, supply-chain integration og cloud computing bliver mere udbredt er der scenarier hvor hændelseslogs fra flere organisationer er til stede for analyser. I princippet er der to indstillinger for tværorganisatorisk process mining.
For det første kan vi tage den kollaborative indstilling, hvor forskellige organisationer arbejder sammen for at håndtere proces instancer. Man kan forestille sig en sådan tværorganisatorisk proces som et puslespil, dvs, den overordnede proces er skåret ud i småstykker og distribueret ud over organisationer, som er nødt til at samarbejde for at samle sagerne (cases). Analyse af hændelsesloggen indenfor en af disse involverede organisationer er ikke tilstrækkeligt. For at identificere end-to-end processer er det nødvendigt at sammenflette hændelseslogs fra forskellige organisationer. Dette er ikke en triviel opgave, da hændelser er nødt til at blive korreleret på tværs af organisatoriske grænser.
For det andet kan vi tage den situation hvor forskellige organisationer essentielt udfører den samme proces og deler erfaringer, viden eller en fælles infrastruktur. Tænk for eksempel på Salesfoce.com. Salgsprocessen fra mange organisationer er sammenlagt og understøttet af Salesforce. På den ene side deler disse organisationer infrastruktur (processer, databaser osv.). På den anden side er de ikke tvunget til at følge en stringent proces model, da systemet kan konfigureres til at understøtte varianter af den samme proces. Som et andet eksempel kan man forestille sig de basale processer, der udføres i hvilken som helst kommune (for eksempel udstede byggetilladelser). Selvom alle kommuner i et land er nødt til at understøtte de samme basale processer, kan der også være forskelle. Det er indlysende, at det er interessant at analysere sådanne varianter mellem forskellige organisationer. Disse organisationer kan lære af hinanden og service udbydere kan forbedre deres services og tilbyde værdi-forøgede services baseret på resultaterne fra tværorganisatorisk process mining. Nye analyse teknikker er nødt til at blive udviklet til begge typer af tværorganisatorisk process mining. Disse teknikker bør også tage hensyn til privatlivs- og sikkerheds-spørgsmål. Organisationer ønsker måske ikke at dele information af konkurrencemæssige grunde eller på grund af manglende tillid. Derfor er det vigtigt at udvikle privatlivs-sikre process mining teknikker.

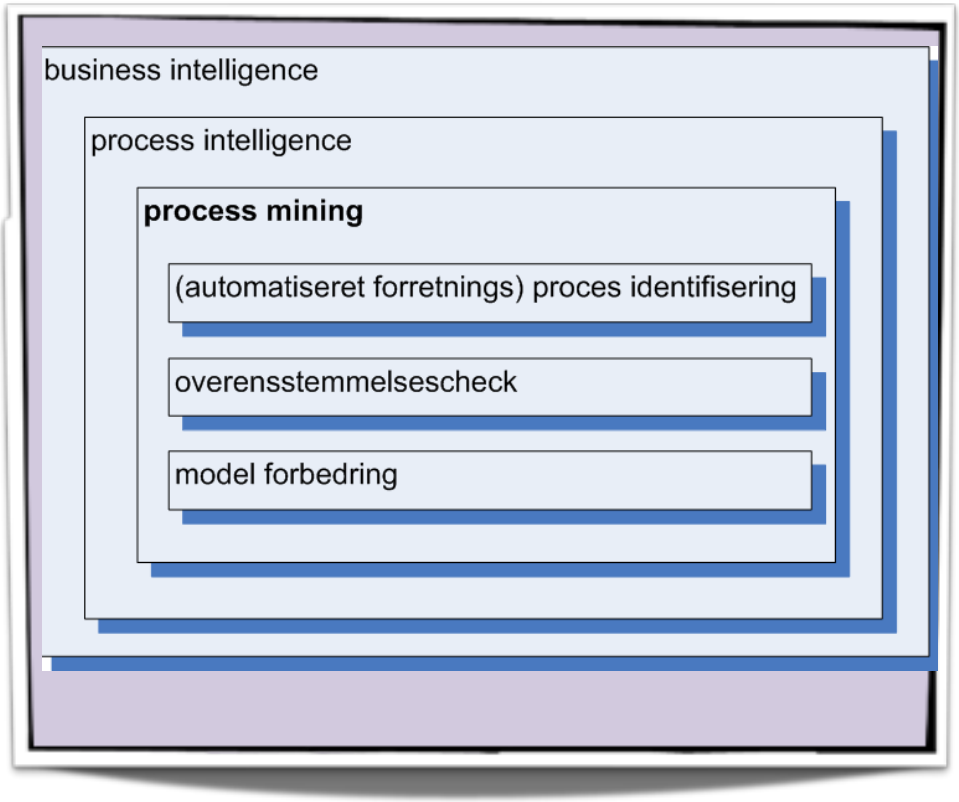
U8: Yde operationel støtte

Initielt har process minings fokus været på at analysere historiske data. Idag er mange data kilder dog opdateret i (næsten) real tid, og tilstrækkelig computer kraft er til rådighed til at analysere hændelser, mens de sker. Derfor bør process mining ikke være begrænset til offline analyser, men kan også bruges til online operationel understøttelse. Der kan identificeres tre operationel understøttelse aktiviteter: *opdag*, *forudse og anbefal*. Det øjeblik hvor en sag (case) afviger fra en forud defineret proces, kan det opdages, og systemet kan generere en advarsel. Ofte vil man gerne straks generere

sådanne advarsler (for at være i stand til at påvirke dem) og ikke på offline måden. Historiske data kan bruges til at bygge forudseende modeller. Disse kan bruges til at vejlede igangværende proces instancer. For eksempel er der muligt at forudse den tilbageværende gennemløbstid for en sag (case). Baseret på sådanne forudsigelser kan man også opbygge anbefalelsessystemer, som foreslår bestemte tiltag for at reducere omkostninger eller for at forkorte gennemløbstiden. Brug af process mining teknikker i sådanne online setup giver yderligere udfordringer i form af computer kraft og data kvalitet.

U9: Kombinere Process Mining med andre analysetyper

Operations management, og i særdeleshed operations research, er en gren af management science som baserer sig kraftigt på modellering. Her benyttes en række forskellige matematiske modeller lige fra linear programming og projekt planlægning til queueing modeller, Markov chains og simulering.
Data mining kan defineres som "analyse af (ofte store) data mængder for at finde uventede sammenhænge og for at summere data på nye måder, som både er forståelige og brugbare for data ejeren".
En bred vifte af teknikker er blevet udviklet: klassificering (for eksempel beslutningstræer), regression, klustering (for eksempel k-means clustering) og mønster genkendelse (for eksempel association rule learning).
Begge områder (operations management og data mining) har værdifulde analyse teknikker. Udfordringen er at kombinere teknikkerne i disse områder med process mining. Tag for eksempel simulering. Process mining teknikker kan bruges til at finde en simuleringsmodel baseret på historiske data. Efterfølgende kan simuleringsmodellen bruges til at give operationel støtte. På grund af den tætte forbindelse mellem hændelseslogs og modellen kan modellen bruges til at afspille historikken, og man kan starte simuleringer fra nuværende tilstand og derved give en "fast forward" knap ind i fremtiden baseret på live data.



På samme måde er det ønskeligt at kombinere process mining med *visual analytics*. Visual analytics kombinerer automatiseret analyse med interaktiv visualiseringer for at få en bedre forståelse for store og komplekse data mængder. Visual analytics gengiver mennerskers utrolige evne til at se mønstre i ustrukturerede data. Ved at kombinere automatiserede process mining teknikker med interaktiv visual analytics, er det muligt at udtrække mere viden fra hændelsesdata.

U10: Forbedre Usability for ikke-eksperter

Et af målene med process mining er at bygge "levende proces modeller". For eksempel proces modeller der benyttes på daglig basis istedet for statiske modeller, der ender i et eller andet arkiv. Nye hændelsesdata kan bruges til at identificere kommende adfærd. Linket mellem hændelsesdata og proces modeller muliggør at nuværende tilstand, og tidligere aktiviteter bæres over i up-to-date modeller.
Hvorved slutbrugere kan interagere med resultaterne fra process mining på en daglig basis.
Sådanne interaktioner er meget værdifulde, men kræver også intuitive bruger interfaces (GUI).
Udfordringen er at skjule de sofistikerede process mining algoritmer bag et brugervenligt interface, som automatisk sætter parametre og foreslår relevante analysetyper.

U11: Forbedre forståelighed for ikke-eksperter

Selv hvis det er let at generere process mining resultater, betyder det ikke at resultaterne rent faktisk er brugbare. Brugeren kan have problemer med at forstå outputtet eller være fristet til at hoppe til forkerte konklusioner. For at undgå sådanne problemer bør resultaterne præsenteres ved brug af en passende repræsentation (se også GP5). Endvidere bør resultaternes troværdighed altid være helt tydelig. Der kan være for lidt data til at retfærdiggøre bestemte konklusioner. Faktisk advarer eksisterende process identificering teknikker typisk ikke om for lav fitness eller om overfitting. De viser altid en model, selv når det er tydeligt at der er for lidt data til at retfærdiggøre nogen konklusioner.

Epilog

The IEEE Task Force on Process Mining søger at (a) promovere process mining, (b) vejlede software udviklere, konsulenter, ledere og slutbrugere i brugen af state-of-the-art teknikker og (c) stimulere forskningen i process mining. Dette manifest fastsætter hovedprincipperne og intensionerne fra task force'n. Efter introduktion af emnet

process mining samler manifestet nogle vejledende principper (sektion 3) og udfordringer (sektion 4). De vejledende principper kan bruges til at undgå oplagte fejltagelser. Listen af udfordringer er tiltænkt forsknings og udviklings tiltag. Begge dele sigter mod at øge process mining modenheden.

Som konklusion, et par ord om terminologi. De følgende termer bruges i process mining branchen: workflow mining, (business) process mining, automated (business) process discovery og (business) process intelligence. Forskellige organisationer bruger forskellige termer for overlappende koncepter. For eksempel promoverer Gartner termen "Automated Business Process Discovery" (ABPD) og Software AG bruger "Process Intelligence" til at referere til deres kontrollerende platform. Termen "workflow mining" ser ud til at være mindre egnet da udarbejdelse af workflow modeller, blot er en af mange anvendelser af process mining

På samme måde indskrænker tilføjelsen af "business" scopet for bestemte anvendelser af process mining. Der er utallige anvendelsesmuligheder for process mining (for eksempel analyse af brugen af high-tech systemer eller analyse af web-sites), hvor denne tilføjelse forekommer upassende. Selvom process discovery er en vigtig del af process mining, er det kun en del af mange brugsscenarier. Overensstemmelse check, forudselser, organisatorisk mining, socialt netværksanalyse osv. er andre brugsscenarier som går videre end process discovery.

Figur 7 relaterer nogle af termerne, der lige er nævnt. Alle teknologier og metoder, som søger at give brugbar information, der kan bruges til til at understøtte beslutninger kan lægges ind under paraplyen Business intelligence (BI). (Business) process intelligence kan ses som en kombination af BI og BPM, da BI teknikker bruges til at analysere og forbedre processer og deres styring. Process mining kan ses som en koncentration af process intelligence, med udgangspunkt i hændelseslogs. (Automated business) process discovery er blot en af tre af de basale process mining typer. Figur 7 er måske lidt misvisende på den måde, at de fleste BI værktøjer ikke har process mining funktionalitet som beskrevet i dette dokument. Termen BI twistes ofte belejligt mod et bestemt værktøj eller metode, dækkende over kun en lille del at det bredere BI spektrum. Der kan være kommercielle grunde til at bruge alternative termer. Nogle leverandører ønsker måske at understrege et bestemt aspekt (fx discovery eller intelligence).

Det er dog bedre at bruge termen "process mining" for disciplinerne, der dækkes af dette manifest.

Ordliste

Aktivitet: Et veldefineret trin i processen. Hændelser kan referere til start, slutning, annullering etc. for en aktivitet for en bestemt proces instans.
Automated Business Process Discovery: se Process Discovery.
Business Intelligence (BI): bred samling af værktøjer og metoder der bruger data til at understøtte beslutninger.
Business Process Intelligence: se Process Intelligence.
Business Process Management (BPM): Disciplinen der kombinerer IT viden og viden om management sciences og bruger begge dele til operationelle forretningsprocesser
Case (sag): se Process Instance.
Concept Drift: Det fænomen hvor processer ofte ændrer sig over tid. Den observerede proces kan langsomt (eller pludseligt) ændre sig grundet sæsonen eller øget konkurrence og dermed komplicere analysen.
Overensstemmelse check: Analysere af om virkeligheden, som lagret i loggen, stemmer overens med modellen og vice versa. Målet er et opdage forskelle og måle deres betydning. Overensstemmelse check er en af de tre typer af process mining.
(Tvær)organisatorisk Process Mining: brugen af process mining teknikker på hændelseslogs der kommer fra forskellige organisationer.
Data Mining: analyse af (ofte store) data mængder for at finde uventede sammenhænge og relationer og summere data på måder, der giver ny indsigt.
Hændelse: en aktion lagret i loggen for eksempel start, slutning eller annullering af en aktivitet for en bestemt proces instans.
Hændelseslog: samling af hændelser brugt som input til process mining. Hændelser behøver ikke at være lagret i separate log filer (for eksempel kan hændelser være spredt ud over forskellige database tabeller).
Fitness: en måling der angiver hvor godt en given model tillader adfærden, der kan ses i hændelsesloggen. En model har en perfekt fitness hvis alle forløb i loggen kan afspilles af modellen fra start til slut.
Generalisering: en måling der angiver hvor godt modellen er i stand til at tillade adfærd, der ikke ses i loggen. En "overfitting" model er ikke i stand til at generalisere nok.
Model forbedring: en af de tre basale typer af process mining. En process model er forbedret eller udvidet ved brug af information udtrukket fra nogle logs. For eksempel kan flaskehalse blive identificeret ved at afspille en hændelseslog på en proces model, mens tidsstempler undersøges.
MXML: et XML-baseret format til at udveksle hændelseslogs. XES erstatter MXML som det nye værktøjs-uafhængige process mining format.
Operationel Support: on-line analyse af hændelsesdata med det formål at monitorere og påvirke kørende proces instanser. Tre operationel support aktiviteter kan identificeres: *opdag* (generer en advarsel hvis den observerede adfærd afviger fra den modellerede adfærd), *forudse* (forudse fremtidig adfærd baseret på historisk adfærd, for eksempel forudse den tilbageværende gennemløbstid), og *anbefal* (foreslå passende aktioner for at realisere en bestemt mål, for eksempel minimere omkostninger).
Præcision: måling der bestemmer om en model forbyder adfærd som er meget anderledes en adfærden set i hændelsesloggen. En model med lav præcision er "underfitting".

Dette manifest blev oprindeligt publiceret i "Business Process Management Workshops 2011, Lecture Notes in Business Information Processing, Vol. 99, Springer-Verlag, 2011, og er siden blevet oversat til flere sprog. Se hjemmesiden for IEEE Task Force on Process Mining: http://www.win.tue.nl/ieeetfpm/ for mere information.

Process Discovery: en af de tre basale typer af process mining. Baseret på en hændelseslog, identificeres en proces model. For eksempel, er α algoritmen i stand til at identificere et Petri net ved at identificere proces mønstre i samlingen af hændelser.
Proces Instans: entiteten der håndteres af processen, der analyseres. Hændelser refererer til proces instanser. Eksempler på proces instanser er kunde ordrer, låne ansøgninger etc.
Process Intelligence: en gren af Business Intelligence der fokuserer på Business Process Management.
Process Mining: teknikker, værktøjer og metoder til at identificere, monitorere og forbedre virkelige processer (dvs. ikke antagede processer) ved at udtrække viden fra almindeligt tilgængelige hændelseslogs i nutidens IT systemer.
Repræsentativ Bias: det valgte sprog til at præsentere og konstruere process mining resultater.
(Simplicity) Enkelhed: en måleenhed der operationaliserer Occam's Razor, dvs den enkleste model, der kan forklare den adfærd, der ses i loggen, er den bedste model. Enkelhed kan kvantificeres på forskellige måder, f.eks, antallet af knudepunkter og buer i modellen.
XES: er en XML-baseret standard for hændelseslogs. Standarden er blevet adopteret af IEEE Task Force on Process Mining som standard udvekslingsformat for hændelseslogs (jf. www.xes-standard.org).

Dansk version af John Hansen, processmining.dk (jha@project2.dk)

Forfattere

Wil van der Aalst
Arya Adriansyah
Ana Karla Alves de Medeiros
Franco Arcieri
Thomas Baier
Tobias Blickle
Jagadeesh Chandra Bose
Peter van den Brand
Ronald Brandtjen
Joos Buijs
Andrea Burattin
Josep Carmona
Malu Castellanos
Jan Claes
Jonathan Cook
Nicola Costantini
Francisco Curbera
Ernesto Damiani
Massimiliano de Leoni
Pavlos Delias
Boudewijn van Dongen
Marlon Dumas
Schahram Dustdar
Dirk Fahland
Diogo R. Ferreira
Walid Gaaloul
Frank van Geffen
Sukriti Goel
Christian Günther
Antonella Guzzo
Paul Harmon
Arthur ter Hofstede
John Hoogland
Jon Espen Ingvaldsen
Koki Kato
Rudolf Kuhn
Akhil Kumar
Marcello La Rosa
Fabrizio Maggi
Donato Malerba
Ronny Mans
Alberto Manuel
Martin McCreesh
Paola Mello
Jan Mendling
Marco Montali
Hamid Motahari Nezhad
Michael zur Muehlen
Jorge Munoz-Gama
Luigi Pontieri
Joel Ribeiro
Anne Rozinat
Hugo Seguel Pérez
Ricardo Seguel Pérez
Marcos Sepúlveda
Jim Sinur
Pnina Soffer
Minseok Song
Alessandro Sperduti
Giovanni Stilo
Casper Stoel
Keith Swenson
Maurizio Talamo
Wei Tan
Chris Turner
Jan Vanthienen
George Varvaressos
Eric Verbeek
Marc Verdonk
Roberto Vigo
Jianmin Wang
Barbara Weber
Matthias Weidlich
Ton Weijters
Lijie Wen
Michael Westergaard
Moe Wynn