Vorlesung

Methodische Grundlagen des Software-Engineering

im Sommersemester 2014

Prof. Dr. Jan Jürjens

TU Dortmund, Fakultät Informatik, Lehrstuhl XIV

Teil 2.0: Einführung: Process-Mining

v. 06.05.2014

2.0 Einführung: Process-Mining

[mit freundlicher Genehmigung basierend auf einem englischen Foliensatz von Prof. Dr. Wil van der Aalst (TU Eindhoven)]

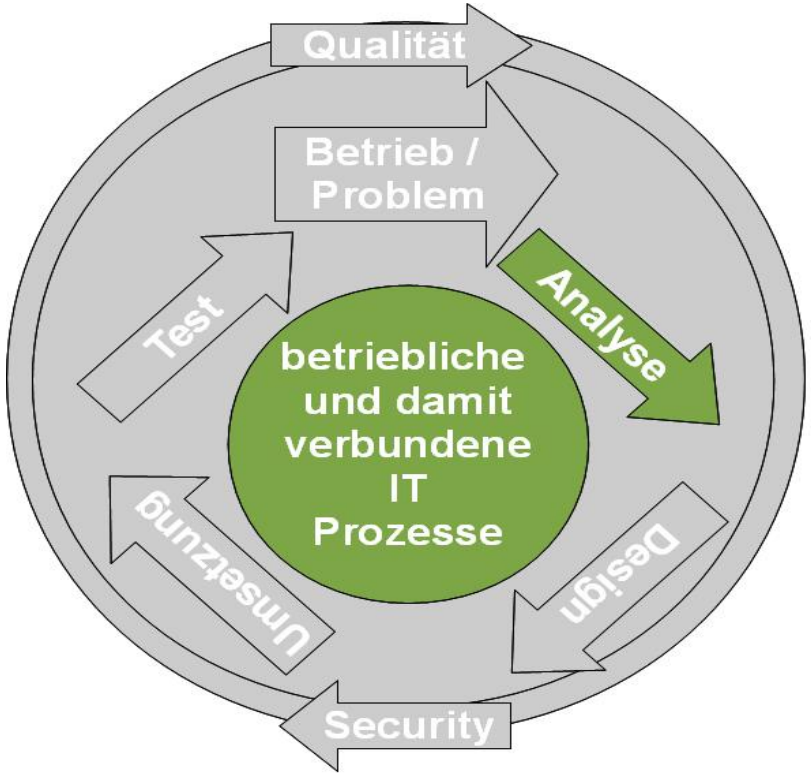
Literatur:

[vdA11] Wil van der Aalst: **Process Mining: Discovery, Conformance and Enhancement of Business Processes**, Springer-Verlag. 2011.
Unibibliothek (6 Exemplare): http://www.ub.tu-dortmund.de/katalog/titel/1332248
(Bei Engpässen kann eine **Kopiervorlage** der relevanten Ausschnitte zur Verfügung gestellt werden.)

- **Kapitel 1**

Einordnung
2.0: Einführung: Process-Mining

- Geschäftsprozessmodellierung
- Process-Mining
  - Einführung: Process-Mining
  - Petrinetze
  - Data-Mining
  - Datenbeschaffung
  - Prozessextraktion
  - Konformanzanalyse
  - Mining: Zusätzliche Perspektiven
  - Betriebsunterstützung
  - Werkzeugunterstützung
  - Analysiere „Lasagne Prozesse“
  - Analysiere „Spaghetti Prozesse“
  - Kartographie und Navigation
  - Epilog
- Modellbasierte Entwicklung sicherer Software



- Letzter Abschnitt 1.4: Von **Modellen** zur **Ausführung** durch **Workflowautomatisierung**.
- Dieses Kapitel 2: Von **Ausführungsdaten** zu **Modellen** (manuelle Erstellung vermeiden; Konformanz überprüfen).
  - Dieser Abschnitt 2.0: Kurze **Einführung** und **Motivation** für die Inhalte von Kapitel 2: **Business Process Mining**.

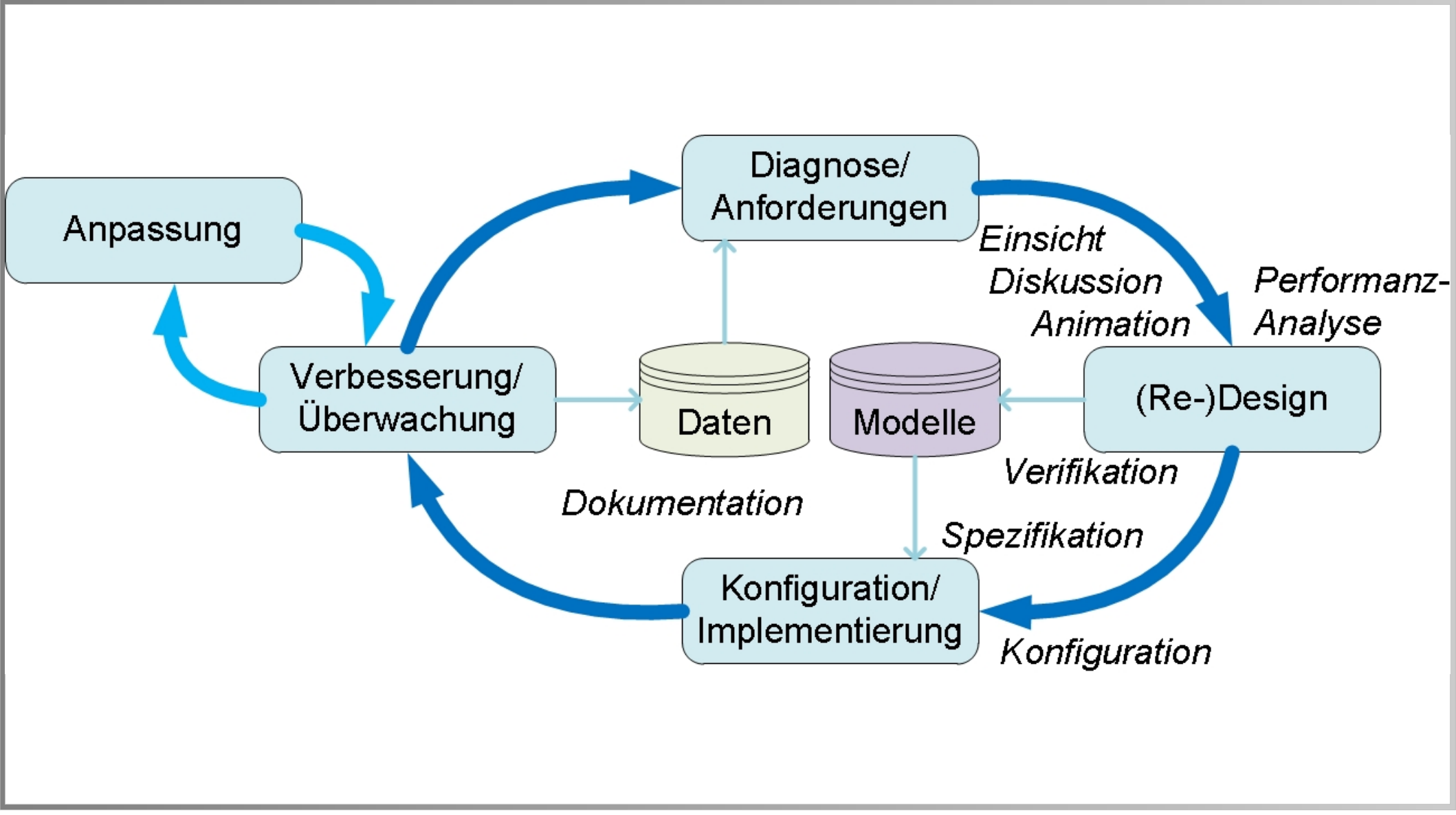
Überblick Process Mining

- **Motivation für Business Process Mining**
- **Business Process Mining:**
  Konformanz, Extraktion, Verbesserung

Traditionelle Verwendung von Geschäftsprozessmodellen

- Einsicht: Prozess von verschiedenen Seiten betrachten.
- Diskussion: Stakeholder können Entscheidungen strukturieren.
- Dokumentation: Personen anleiten, Zertifizierungen erhalten (z.B. ISO 9000 Qualitätsmanagement).
- Verifikation: Prozessmodelle analysieren: Fehler in System oder Prozeduren finden (z.B. Deadlocks).
- Performance-Analyse: Simulation => z.B. beeinflussende Faktoren für Antwortzeiten, Service Level etc. verstehen.
- Animation: End-Nutzer können Szenarios “durchspielen”; Feedback an Designer geben.
- Spezifikation: Process-aware Information System (PAIS) vor Implementierung modellieren (“Vertrag zwischen Entwicklern und End-Nutzer / Management“).
- Konfiguration: Modelle als Konfiguration von Systemen.

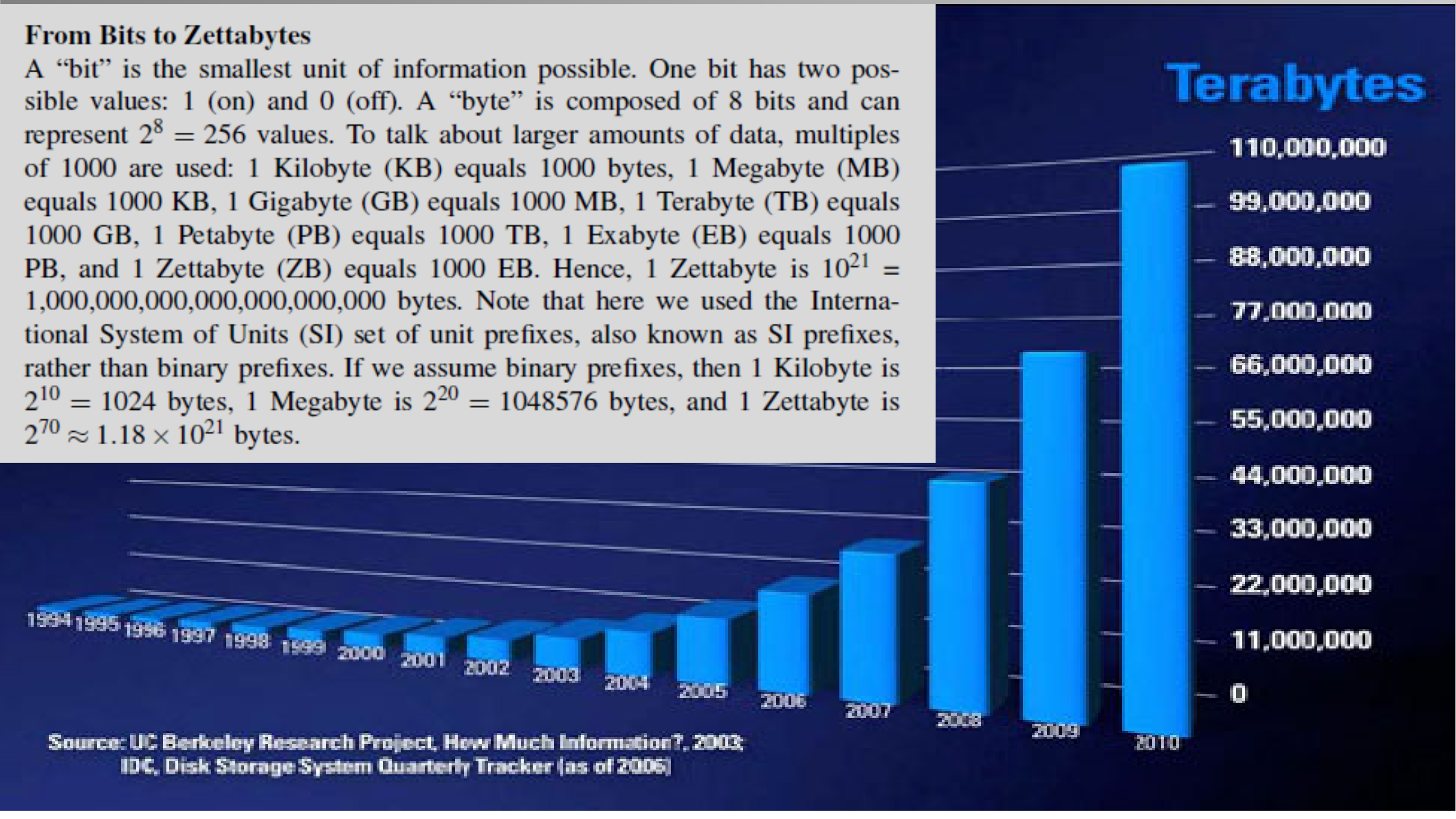
BPM-Lebenszyklus: klassischer Gebrauch von Prozessmodellen



Neue Entwicklung: Datenexplosion

From Bits to Zettabytes

A "bit" is the smallest unit of information possible. One bit has two possible values: 1 (on) and 0 (off). A "byte" is composed of 8 bits and can represent $2^8 = 256$ values. To talk about larger amounts of data, multiples of 1000 are used: 1 Kilobyte (KB) equals 1000 bytes, 1 Megabyte (MB) equals 1000 KB, 1 Gigabyte (GB) equals 1000 MB, 1 Terabyte (TB) equals 1000 GB, 1 Petabyte (PB) equals 1000 TB, 1 Exabyte (EB) equals 1000 PB, and 1 Zettabyte (ZB) equals 1000 EB. Hence, 1 Zettabyte is 10^{21} = 1,000,000,000,000,000,000,000 bytes. Note that here we used the International System of Units (SI) set of unit prefixes, also known as SI prefixes, rather than binary prefixes. If we assume binary prefixes, then 1 Kilobyte is $2^{10} = 1024$ bytes, 1 Megabyte is $2^{20} = 1048576$ bytes, and 1 Zettabyte is $2^{70} \approx 1.18 \times 10^{21}$ bytes.



Real-Daten mit Prozess-Modellen bewältigen

Datenflut bewältigen mittels GP-Modellierung: **GP-Modellierung** bei **Prozessausführung** optimal nutzen (z.B. **Prozessoptimierung**).

Relevante Ansätze:

- Business Process Management (BPM)
- Business Intelligence (BI)
- Online Analytical Processing (OLAP)
- Business Activity Monitoring (BAM)
- Complex Event Processing (CEP)
- Corporate Performance Management (CPM)
- Visual Analytics (VA)
- Predictive Analytics (PA)
- Continuous Process Improvement (CPI)
- Total Quality Management (TQM)
- Six Sigma

Beispiel: Six Sigma

Motorola, frühe 1980er. „DMAIC“-Ansatz:

- Definiere (**D**efine) Problem und Ziel,
- Messe (**M**easure) Hauptindikatoren für Leistung; sammele Daten,
- Analysiere (**A**nalyze) Daten: Zusammenhang zwischen Ursache und Wirkung ermitteln und verifizieren,
- Verbessere (**I**mprove) aktuellen Prozess basierend auf Analyse
- Kontrolliere (**C**ontrol) Prozess: Zielabweichung minimieren.

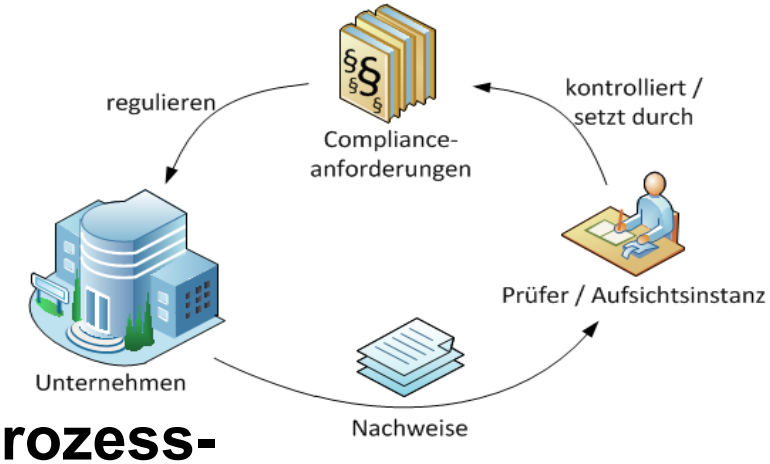
Aktuelle Herausforderung: GP-Modellierung für Compliance

Steigende Anzahl von **Regulierungen**

- z.B. **Finanzen**: Solvency II, Basel III; **Gesundheit**: Medizinproduktegesetz (MPG); **Pharma**: Arzneimittelmarktneuordnungsgesetz (AMNOG)

Steigende Komplexität des **Compliance-Nachweises:**

- **Viele Facetten**: Anforderungen an Geschäftsprozesse (Design + Ausführung), IT-Infrastruktur (+ deren Prozesse), deren Abhängigkeiten…
- **Wechselseitige Abhängigkeiten** von Vorgaben
- **Aggregation** von Vorgaben bei komplexen IT-Systemen
- **Cloud-Computing** → besondere Anforderungen



=> **Nachweis** der **Compliance** mit **Geschäftsprozess-modellierung** und **-überwachung** erleichtern.

Nachteile manueller Prozessmodellierung

Was sind Nachteile von manueller Prozessmodellierung ?

Nachteile manueller Prozessmodellierung

Was sind Nachteile von manueller Prozessmodellierung ?

(1) "Handgemachte" Modelle **unterscheiden** sich oft von Realität: idealisierte Sicht auf Prozess ("Papier-Tiger").

- (Nur) **ausführbare** Modelle können bestimmte Arbeitsweise tatsächlich **erzwingen** (vgl. Compliance-Nachweis).

(2) **Manuelle** Erstellung der Modelle oft **aufwendig**.
=> **Prozessänderungen** im Modell oft **nicht nachgezogen**.

Wie kann ich diese Nachteile überwinden ?

Von Prozessmodellierung zu Process-Mining

Was sind Nachteile von manueller Prozessmodellierung ?

(1) "Handgemachte" Modelle **unterscheiden** sich oft von Realität: idealisierte Sicht auf Prozess ("Papier-Tiger").

- (Nur) **ausführbare** Modelle können bestimmte Arbeitsweise tatsächlich **erzwingen** (vgl. Compliance-Nachweis).

(2) **Manuelle** Erstellung der Modelle oft **aufwendig**.
=> **Prozessänderungen** im Modell oft **nicht nachgezogen**.

Wie kann ich diese Nachteile überwinden (unter Verwendung des Laufzeitverhaltens in Form von generierten Logdaten) ?

(1) => **Konformanz** Laufzeitverhalten <-> Prozessmodell **überprüfen** !

(1) & (2) => **Prozessmodell** aus Laufzeitverhalten **extrahieren** !?

- Erleichtert durch heutige große Menge an Event-Daten.

→ **Process-Mining** !

Process-Mining

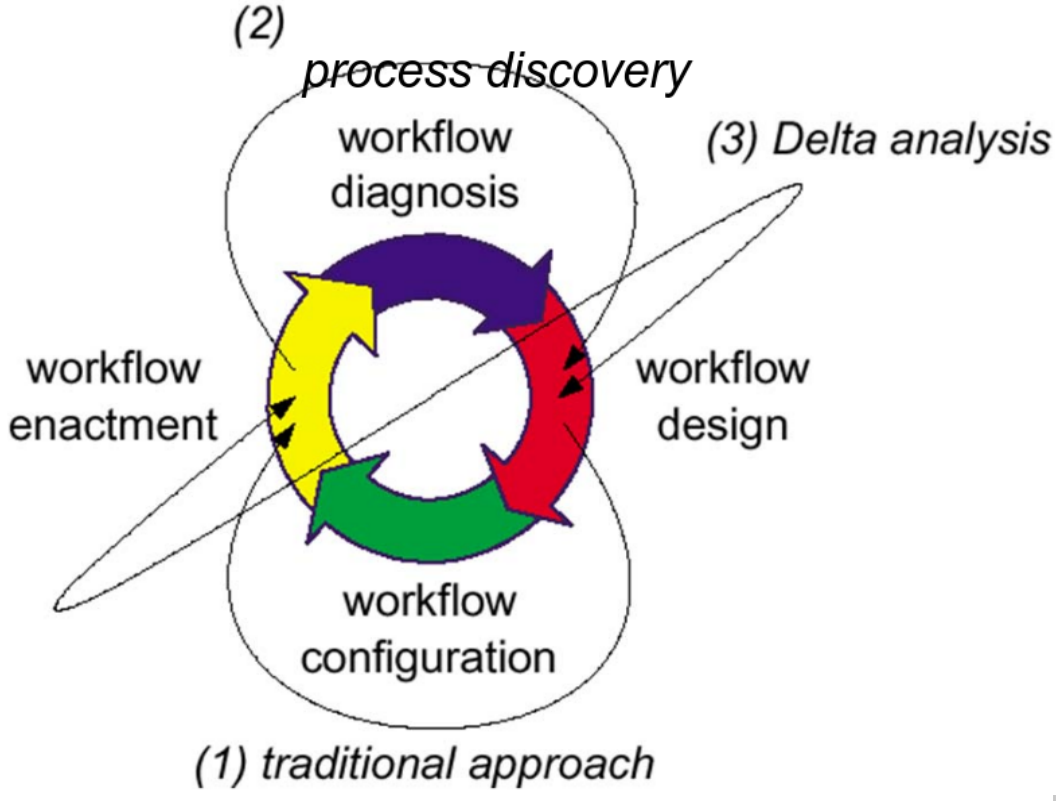
- (Teil-)automatisierte **Extraktion** von **GP-Modellen** aus **Laufzeitdaten** (z.B. Log-Daten).
- **Extraktion** von Modellen basierend auf Fakten.
- Ziel **nicht** Erzeugung eines einzelnen Modells des Prozesses.
- **Sondern**: verschiedene **Sichten** auf gleiche Realität auf verschiedenen **Abstraktionsebenen**. Beispiel: nur **häufigstes** Verhalten betrachten, für einfaches Modell ("**80%-Modell**").
- Auch **gesamtes** Verhalten betrachtbar ("**100%-Modell**", deckt alle beobachteten Fälle ab).

Workflow-Lebenszyklus

Klassischer „**Lebenszyklus**“ von Workflows:

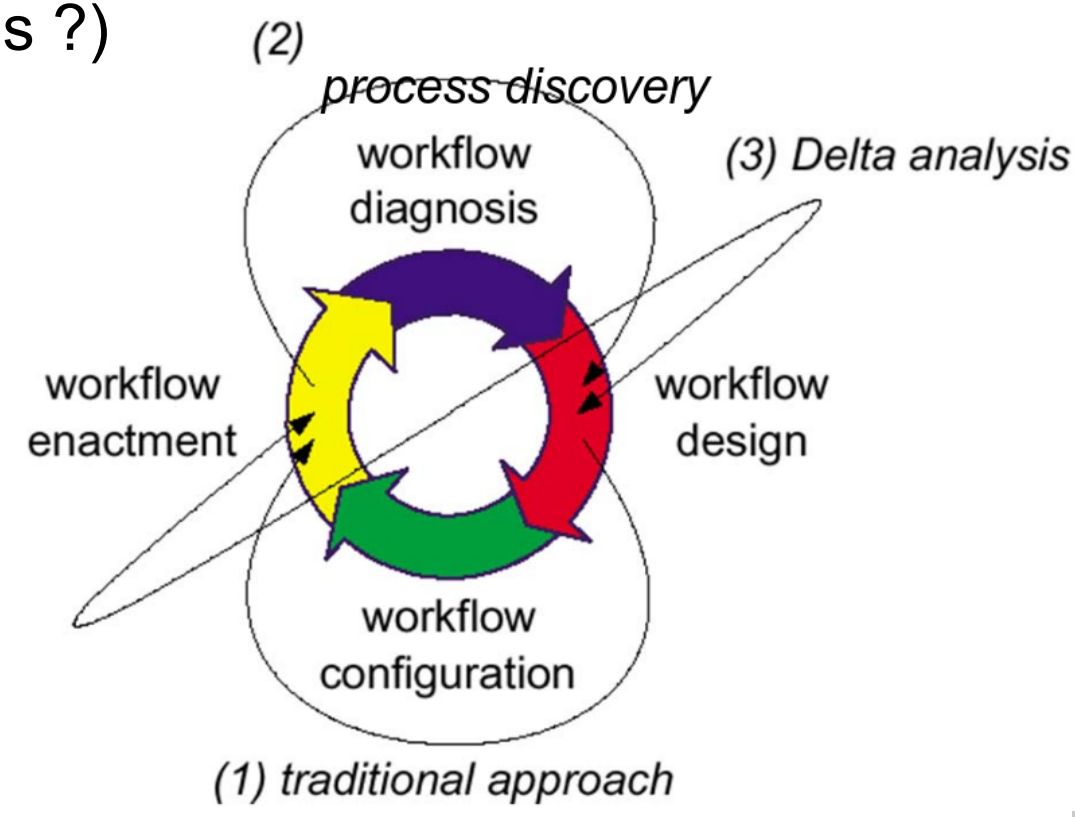
- Workflow **design**: Entwurf des Workflows
- Workflow **configuration**: Konfiguration des Workflows
- Workflow **enactment**:
  Ausführung des Workflows
- Workflow **diagnosis**:
  Diagnose des
  ausgeführten Workflows

Klassisches Vorgehen (1):
Workflow-**Ausführung** auf
Basis des Workflow-**Designs**.



Ansätze auf Basis von **Process-Mining**:

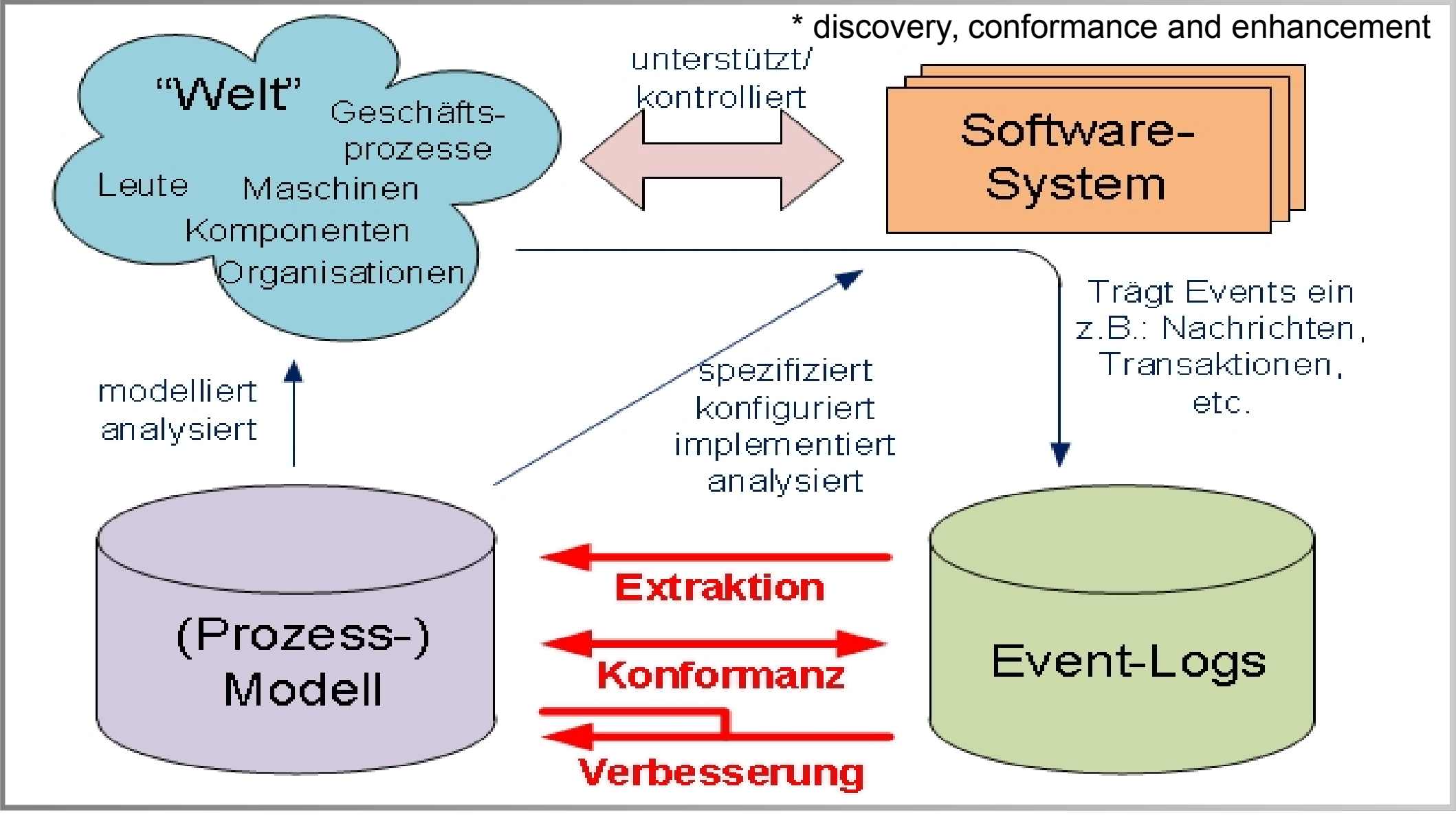
- **Process Discovery (2)**: **Extraktion** des Workflow-**Designs** für **ausgeführten Workflow**.
  (Wie sieht Prozess (wirklich) aus ?)
- **Delta Analysis (3)**: **Vergleich** Workflow-**Design** mit **ausgeführtem Workflow**.
  (Passiert das, was spezifiziert wurde ?)



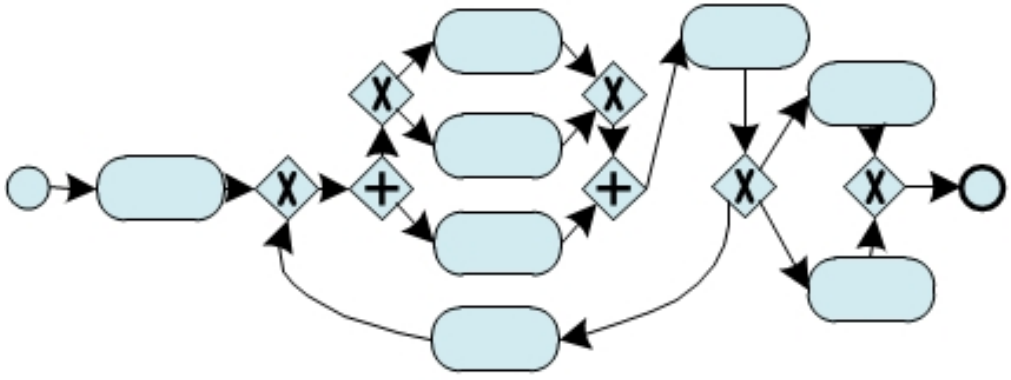
Überblick Process Mining

- Motivation für Business Process Mining
- **Business Process Mining:**
  Konformanz, Extraktion, Verbesserung

Prozess-Modell vs. Event-Logs: Konformanz, Extraktion, Verbesserung*

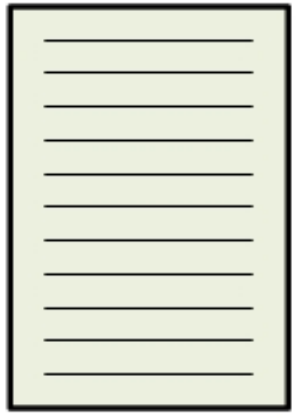


Konformanz zwischen Prozessmodell und Event-Log ?



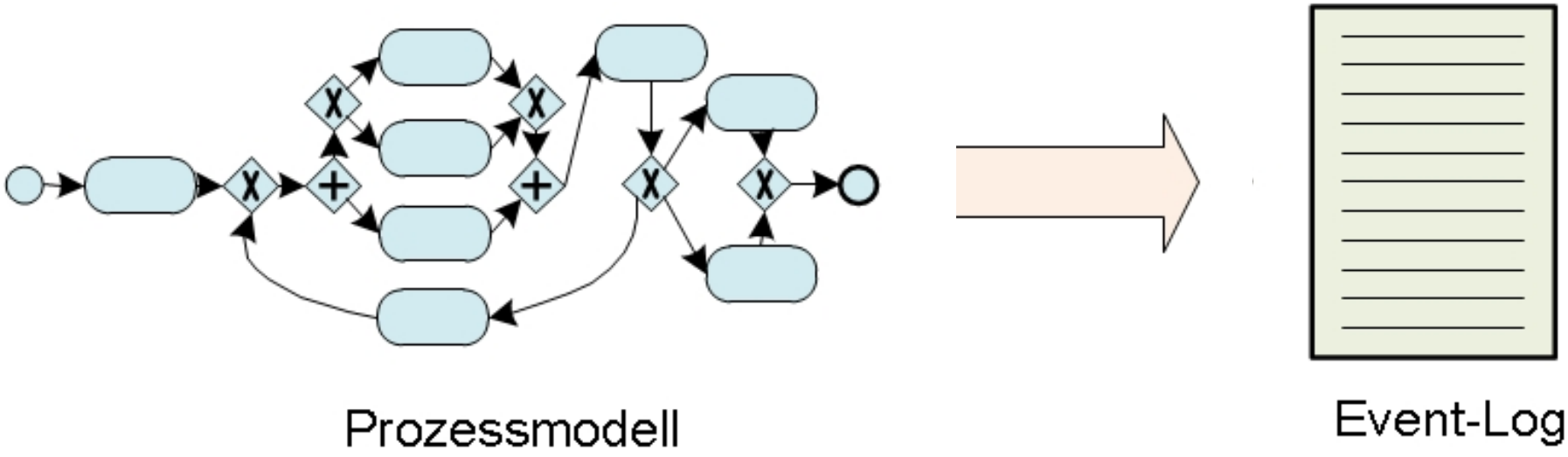
Prozessmodell

?



Event-Log

Konformanz mittels “Play-Out“



Prozessmodell

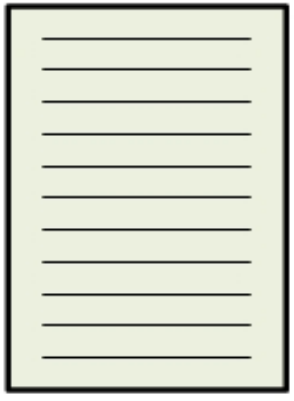
Event-Log

Aus Prozessmodell Menge der möglichen Event-Logs generieren.

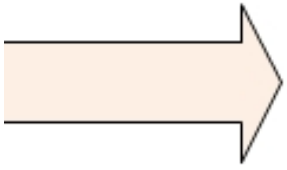
Beobachtetes Event-Log ist konformant zu Prozessmodell, wenn es in dieser Menge enthalten ist.

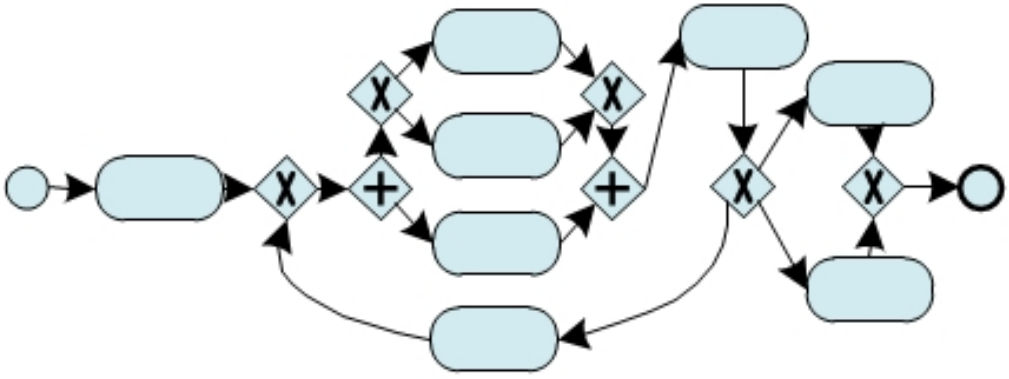
(Geht nur, wenn Menge der möglichen Event-Logs endlich. Ansonsten Konformanz „online“ überprüfen, d.h. iterativ jeweils die Menge der möglichen nächsten Prozessschritte generieren und mit dem nächsten Log-Datum vergleichen.)

Extraktion durch “Play-In“



Event-Log



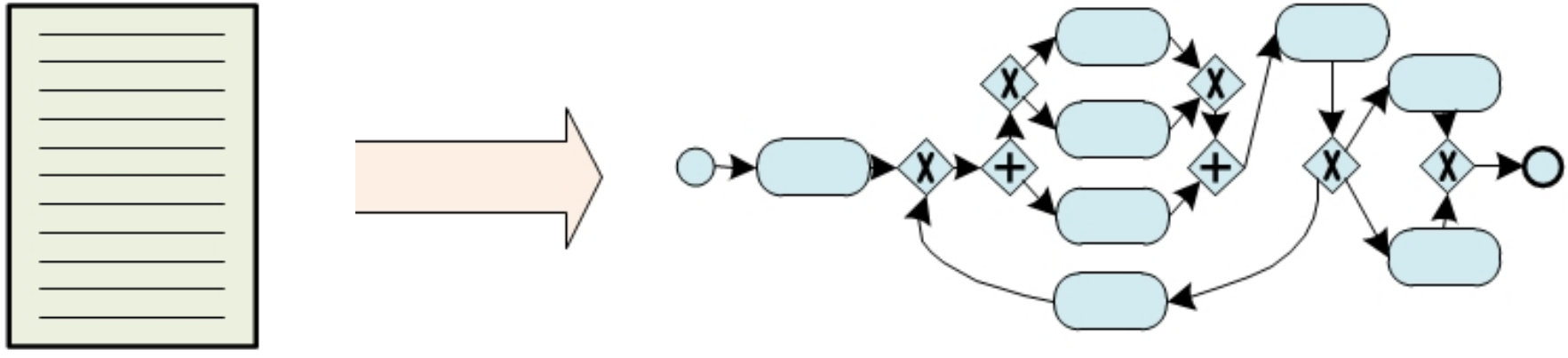


Prozessmodell

Aus gegebener Menge Event-Logs automatisch zugehöriges Prozessmodell generieren.

Was genau heisst „**zugehöriges**“ ?

Extraktion durch “Play-In“



Event-Log

Prozessmodell

Aus gegebener Menge Event-Logs automatisch zugehöriges Prozessmodell generieren.

Was genau heisst „**zugehöriges**“ ?

- Ursprüngliche Event-Logs **konform** zu generiertem Prozessmodell: Häufig gefordert (aber nicht immer: „80%-Modell“ - nur häufigstes Verhalten in Modell abbilden, z.B. zur Modellvereinfachung).
- Gibt mehr als ein solches Prozessmodell: Kann z.B. zusätzliche Verzweigungen hinzufügen. Wäre aber nicht sinnvoll, also oft möglichst „**präzises**“ oder „**einfaches**“ Modell gesucht. Nicht-trivial zu defineren; ggf. **trade-offs**.

Prozessextraktion Beispiel: Startpunkt: Event-Log

case id	event id	properties				
		timestamp	activity	resource	cost	...
1	35654423	30-12-2010:11.02	register request	Pete	50	...
	35654424	31-12-2010:10.06	examine thoroughly	Sue	400	...
	35654425	05-01-2011:15.12	check ticket	Mike	100	...
	356					
	356					
2	356					
	356					
	356					
	356					
3	356					
	356					
	356					
	356					
	356					
	356					
	356					
	356					
4	356					
	356					
	356					
	356					
	356					
5	356					
	356					
	356					
	356					
	356					
	356					
	356					
	356					
	356					
	356					
	356					
	356					
6	35654873	06-01-2011:16.06	examine casually	Ellen	400	...
	35654874	07-01-2011:16.22	check ticket	Mike	100	...
	35654875	07-01-2011:16.52	decide	Sara	200	...
	35654877	16-01-2011:11.47	pay compensation	Mike	200	...
...	...	...	...	...	...	...

Fall ID	Event ID	Eigenschaften				
		Zeitstempel	Aktivität	Ressource	Kosten	...
1	35654423	30-12-2010:11.02	Registrierung anfragen	Pete	50	...
	35654424	31-12-2010:10.06	gründlich überprüfen	Sue	400	...
	35654425	05-01-2011:15.12	Ticket überprüfen	Mike	100	...
	35654426	06-01-2011:11.18	entscheiden	Sara	200	...
	35654427	07-01-2011:14.24	Anfrage ablehnen	Pete	200	...
2	35654483	30-12-2010:11.32	Registrierung anfragen	Mike	50	...
	35654485	30-12-2010:12.12	Ticket überprüfen	Mike	100	...
	35654487	30-12-2010:14.16	normal überprüfen	Pete	400	...
	35654488	05-01-2011:11.22	entscheiden	Sara	200	...
	35654489	08-01-2011:12.05	Entschädigung bezahlen	Ellen	200	...

Speicherformate: XES, MXML, SA-MXML, CSV, etc.

Vereinfachter Event-Log

Fall ID	Event ID	Eigenschaften			
		Zeitstempel	Aktivität	Ressource	Kosten
1	35654423	30-12-2010:11.02	Registrierung anfragen	Pete	50
	35654424	31-12-2010:10.06	gründlich überprüfen	Sue	400
	35654425	05-01-2011:15.12	Ticket überprüfen	Mike	100
	35654426	06-01-2011:11.18	entscheiden	Sara	200
	35654427	07-01-2011:14.24	Anfrage ablehnen	Pete	200
2	35654483	30-12-2010:11.32	Registrierung anfragen	Mike	50
	35654485	30-12-2010:12.12	Ticket überprüfen	Mike	100
	35654487	30-12-2010:14.16	normal überprüfen	Pete	400
	35654488	05-01-2011:11.22	entscheiden	Sara	200
	35654489	08-01-2011:12.05	Entschädigung bezahlen	Ellen	200

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

a = Registrierung anfragen (register request),
b = gründlich überprüfen (examine thoroughly),
c = normal überprüfen (examine casually),
d = Ticket überprüfen (check ticket),
e = entscheiden (decide),
f = Anfrage neu einleiten (reinitiate request),
g = Entschädigung bezahlen (pay compensation),
h = Anfrage ablehnen (reject request)

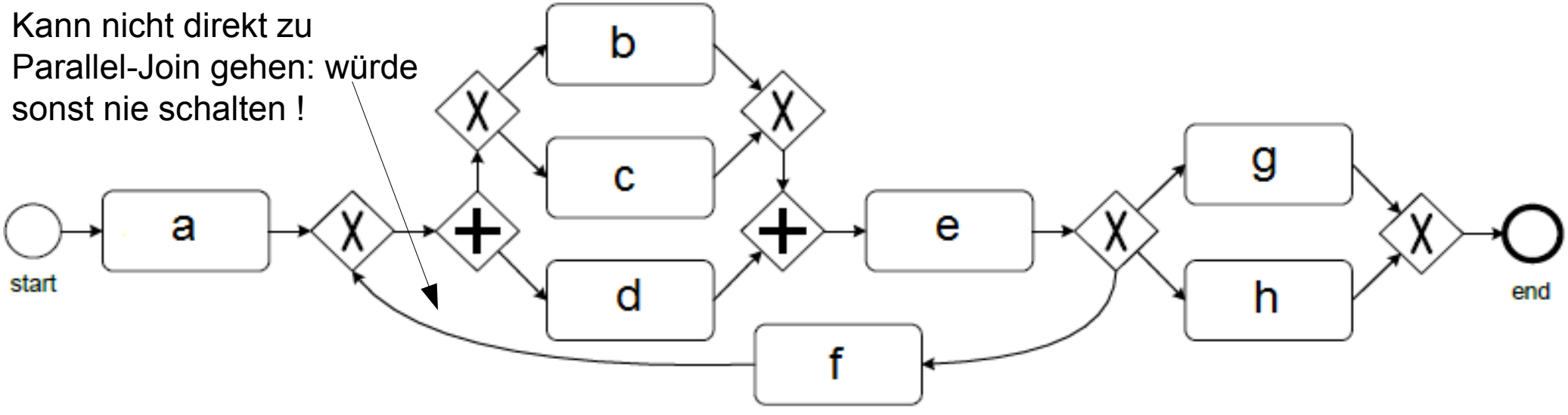
Prozess-Discovery: BPMN-Modell

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

Prozess-Discovery: BPMN-Modell

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

Kann nicht direkt zu Parallel-Join gehen: würde sonst nie schalten !



Mit Process-Mining extrahierte Perspektiven

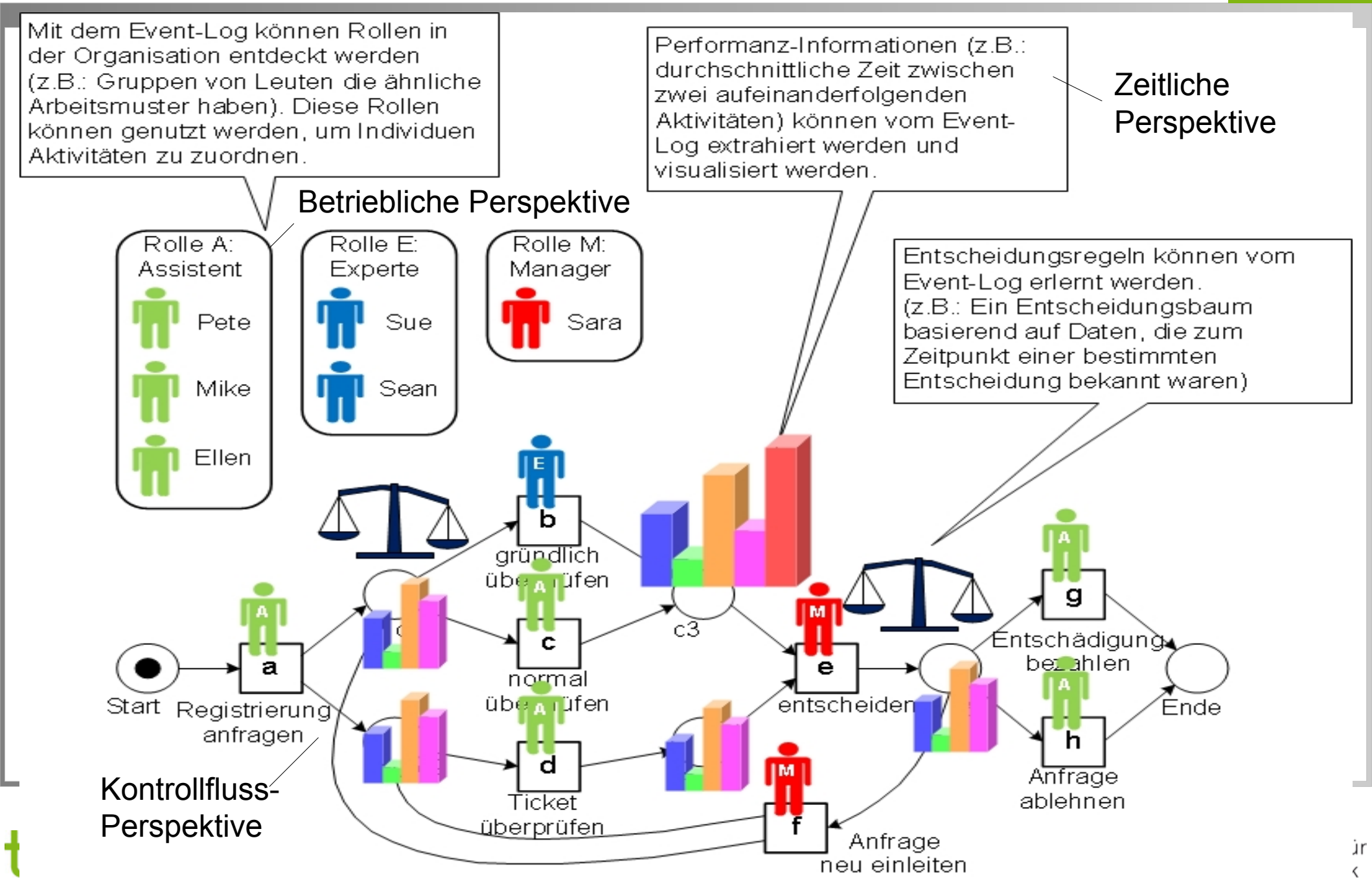
In klassischen Prozessmodellen enthalten (hier aus Logdaten extrahiert):

- Kontrollfluss-Perspektive: Reihenfolge der **Aktivitäten**.
- Betriebliche Perspektive: Informationen über **Ressourcen**: Welche Akteure (z.B.: Systeme) involviert, in welcher Beziehung.

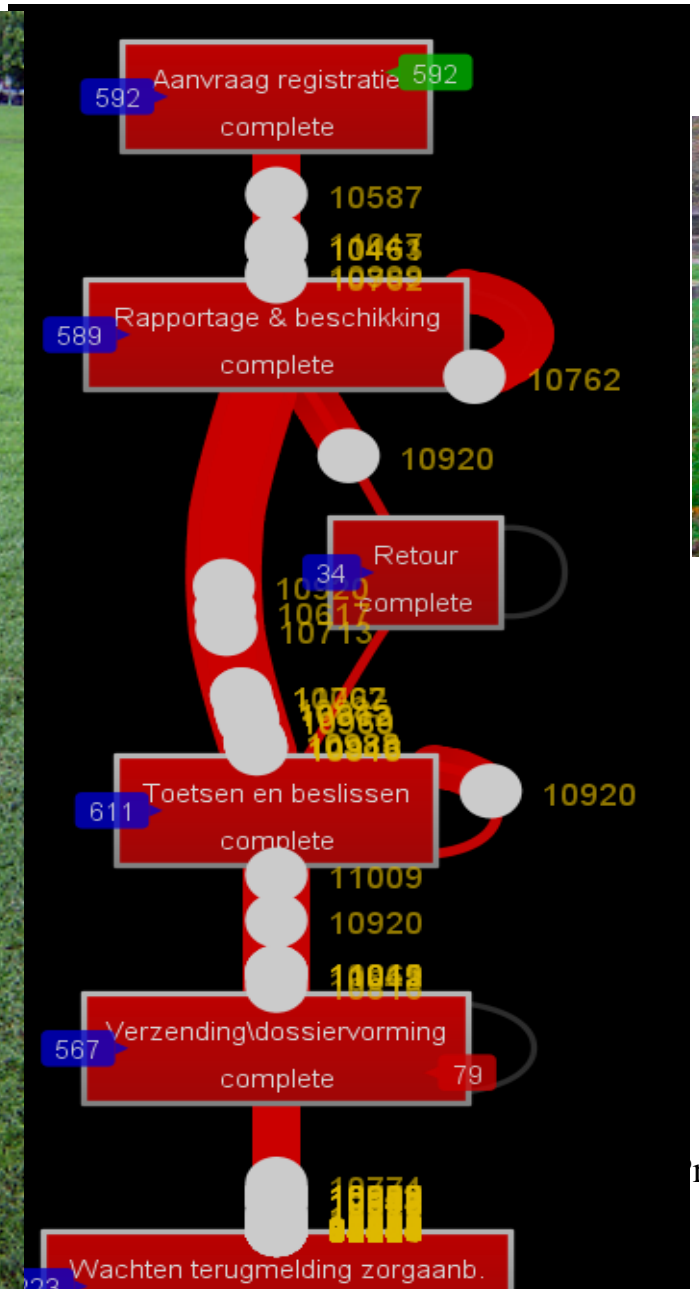
Weitere Informationen über klassische Prozessmodelle hinaus:

- Fall-Perspektive: Eigenschaften von **Fällen**. Z.B. über Werte der zugehörigen Daten-Elemente charakterisierbar.
- Zeitliche Perspektive: Timing und Frequenz von **Ereignissen**.

Process-Mining: Perspektiven

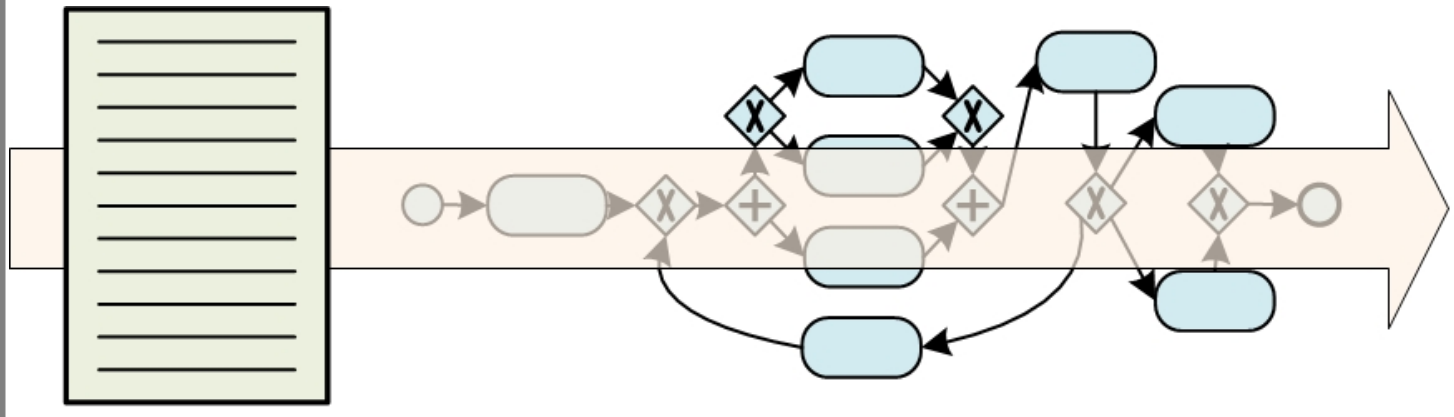


Frequenz von Ereignissen: Wunschpfade in Prozess-Modellen



Spezialfall: Non-konformanz aufdecken

Verbesserung durch Replay



Event-Log

Prozessmodell

- Erweitertes Modell zeigt: Zeiten, Frequenzen, etc.
- Diagnosen
- Vorhersagen
- Empfehlungen

Replay: Ausführungsfolgen aus extrahiertem Modell generieren und mit ursprünglichem Event-Log vergleichen.

- Grad der **Konformanz** überprüfen (bis 80%-Modellen).
- Modelle daraufhin entsprechend **nachjustieren**.
- **Voraussagende** Modelle konstruieren.
- **Betriebsunterstützung** (Vorhersagen, Vorschläge, etc.).

Zusammenfassung
2.0: Einführung: Process-Mining

Dieser Abschnitt: Kurze Einführung und Motivation für die Inhalte von Kapitel 2: Business Process Mining.

- **Motivation:** Traditionelle Verwendung von Prozessmodellierung vs. aktuelle Herausforderungen (Datenexplosion, Compliance).
- **Einführung:** Business Process Mining: Von Ausführungsdaten zu Modellen (manuelle Erstellung vermeiden; Konformanz überprüfen).
  - Konformanz mittels Play-Out
  - Extraktion mittels Play-In
  - Verbesserung mittels Re-Play

Hausaufgabe

Anschauen:

- Informelle Einführung in Process Mining + Verweis auf weitere Informationen (7 min)

http://www.promtools.org/pmtv/movies/pmtv01.mov

(verlinkt von Vorlesungsseite, Abschnitt Vorlesungsfolien)

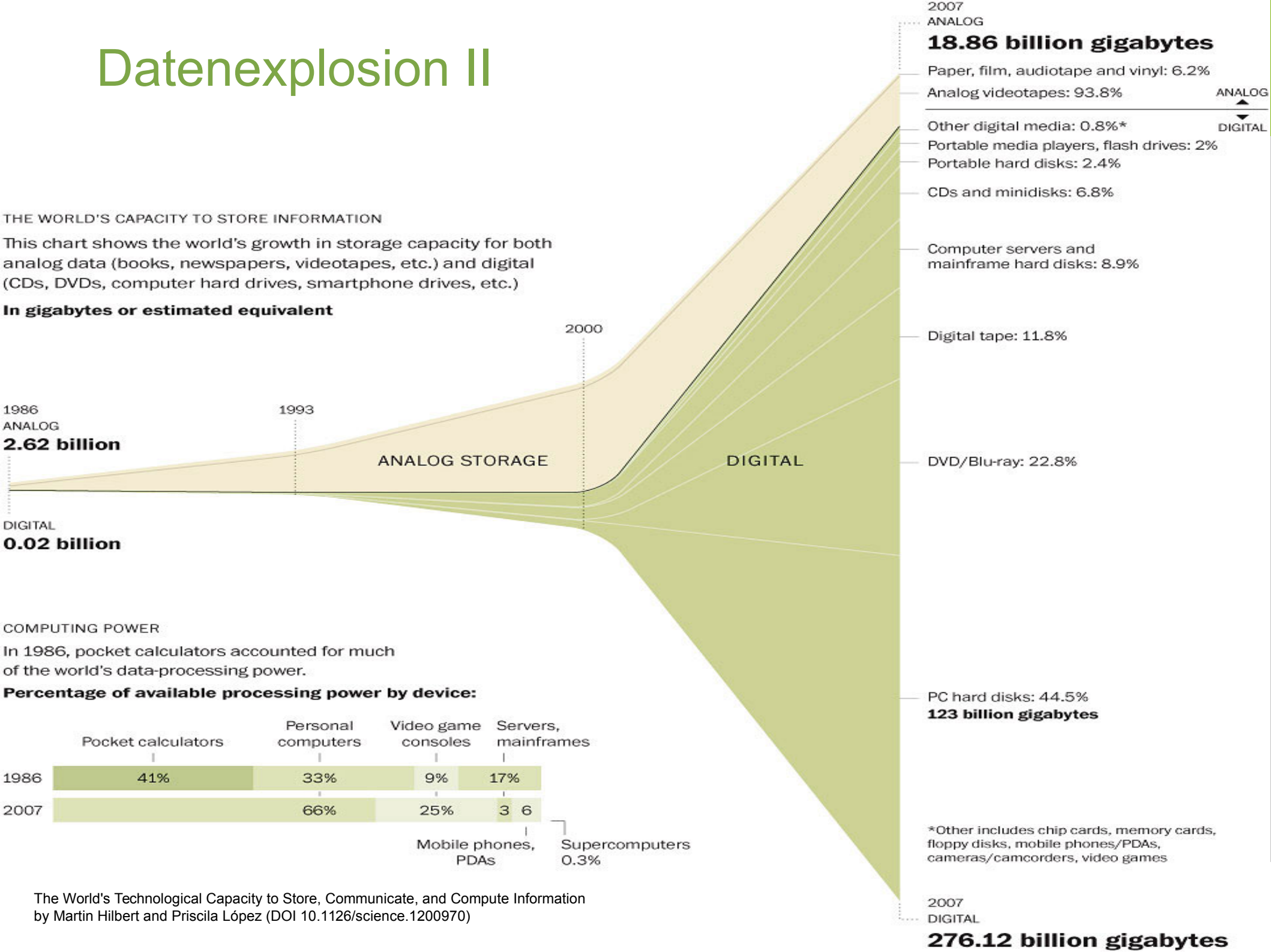
Ausblick Kapitel 2: Process Mining

- Einführung: Process-Mining
- Petrinetze
- Data-Mining
- Datenbeschaffung
- Prozessextraktion
- Konformanzanalyse
- Mining: Zusätzliche Perspektiven
- Betriebsunterstützung
- Werkzeugunterstützung
- Analysiere „Lasagne Prozesse“
- Analysiere „Spaghetti Prozesse“
- Kartographie und Navigation
- Epilog

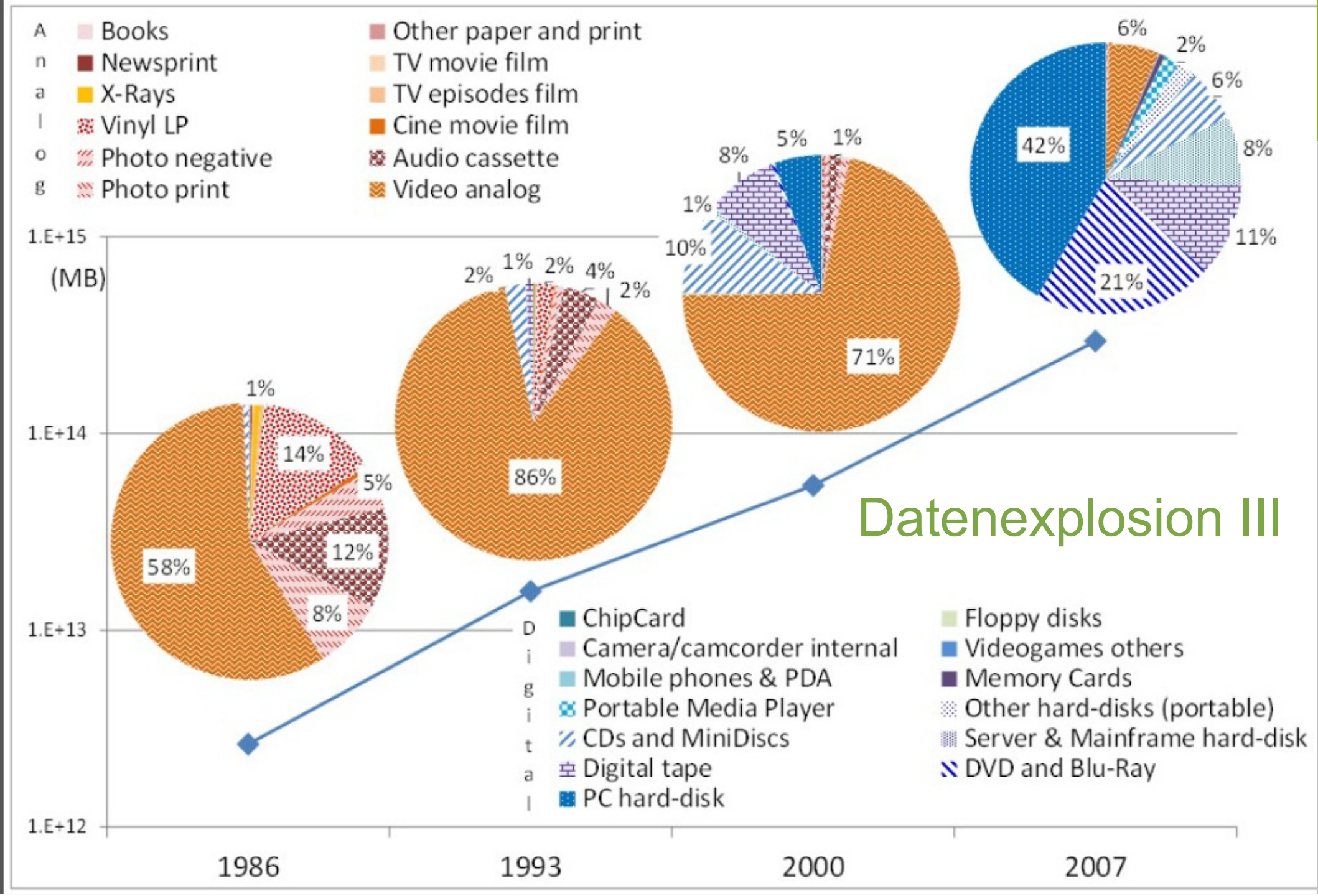
Nächster Abschnitt: Einführung bzw. Wiederholung **Petrinetze** (Grundlage für Process Mining Ansatz).

Anhang (weitere Informationen und Beispiele)

Datenexplosion II



The World's Technological Capacity to Store, Communicate, and Compute Information
by Martin Hilbert and Priscila López (DOI 10.1126/science.1200970)


Analog
Books
Newsprint
X-Rays
Vinyl LP
Photo negative
Photo print
Other paper and print
TV movie film
TV episodes film
Cine movie film
Audio cassette
Video analog
1.E+15
(MB)
1.E+14
1.E+13
1.E+12
1986
1993
2000
2007
1%
14%
5%
12%
8%
58%
2%
1%
2%
4%
2%
86%
8%
5%
1%
1%
10%
71%
6%
2%
6%
8%
11%
21%
42%
Digital
ChipCard
Camera/camcorder internal
Mobile phones & PDA
Portable Media Player
CDs and MiniDiscs
Digital tape
PC hard-disk
Floppy disks
Videogames others
Memory Cards
Other hard-disks (portable)
Server & Mainframe hard-disk
DVD and Blu-Ray
Datenexplosion III

Rolle der Modelle

- Mit **Modellen**: **Prozesse erörtern** (**Re-Design**) und **Entscheidungen** in Prozessen treffen (**Planung** und **Kontrolle**).
- Prozessmodell für Diskussion über:
  - **Verantwortlichkeiten**,
  - **Compliance-Analyse**,
  - **Leistungsvorhersage** durch Simulationen und
  - **Konfiguration** von Workflowmanagement-Systemen.

Einschränkungen Modellbasierter Analyse

Modell:

- **Idealisierte** Version der Realität.
- **Menschliches Verhalten** nicht adäquat darstellbar.
- Oft falsche **Abstraktionsebene**.
- **Verifikation** und **Performanzanalyse** braucht **hochqualitative** Modelle.
- Bei zu großem Unterschied **Modell – Wirklichkeit**: modellbasierte Analyse sinnlos.
- Oft fehlt **Abgleich**: handgemachte Modelle – Wirklichkeit.
  - Geschäftsprozessmodellierung nicht vorhanden / nicht vollständig bzw. aktuell.
  - Wird GP-Dokumentation in der Praxis befolgt ?

Process-Mining: Modell vs. Daten

Process-Mining adressiert die o.g. Probleme:

- **Direkte Verbindung** zwischen **Modell** und **Daten** eines **aktuellen Prozess-Ereignisses**.
- **Prozessextraktions**-Techniken: Sicht aus **verschiedenen Perspektiven** / **Abstraktionsebenen**.