LEHRSTUHL 14
SOFTWARE ENGINEERING

Vorlesung

Methodische Grundlagen des Software-Engineering

im Sommersemester 2013

Prof. Dr. Jan Jürjens

TU Dortmund, Fakultät Informatik, Lehrstuhl XIV

Teil 2.0: Einführung: Process-Mining

v. 29.04.2013

2.0 Einführung: Process-Mining

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

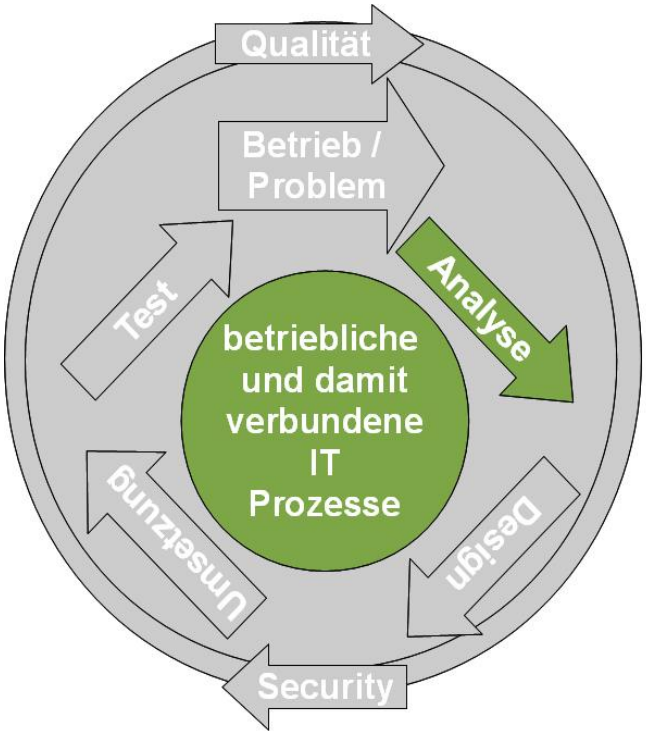
Literatur:

[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

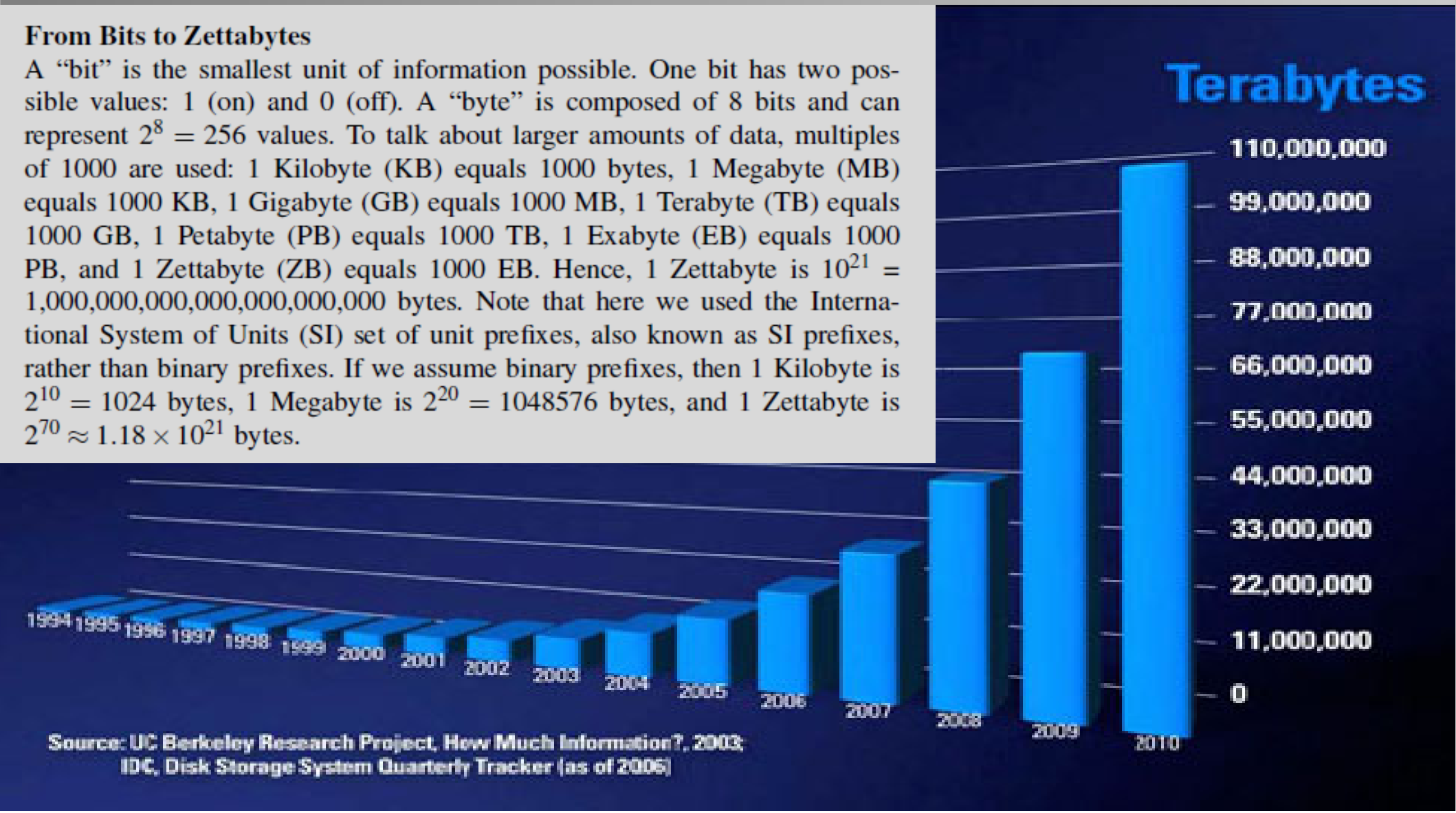
- Geschäftsprozessmodellierung
- Business Process Mining
  - Einführung: Process Mining
  - Petrinetze
  - Prozessmodellierung und Analyse
  - Data-Mining
  - Datenbeschaffung
  - Prozessextraktion
  - Fortgeschrittene Prozessextraktion
- Modellbasierte Softwareentwicklung
- Modellbasierte Entwicklung sicherer Software



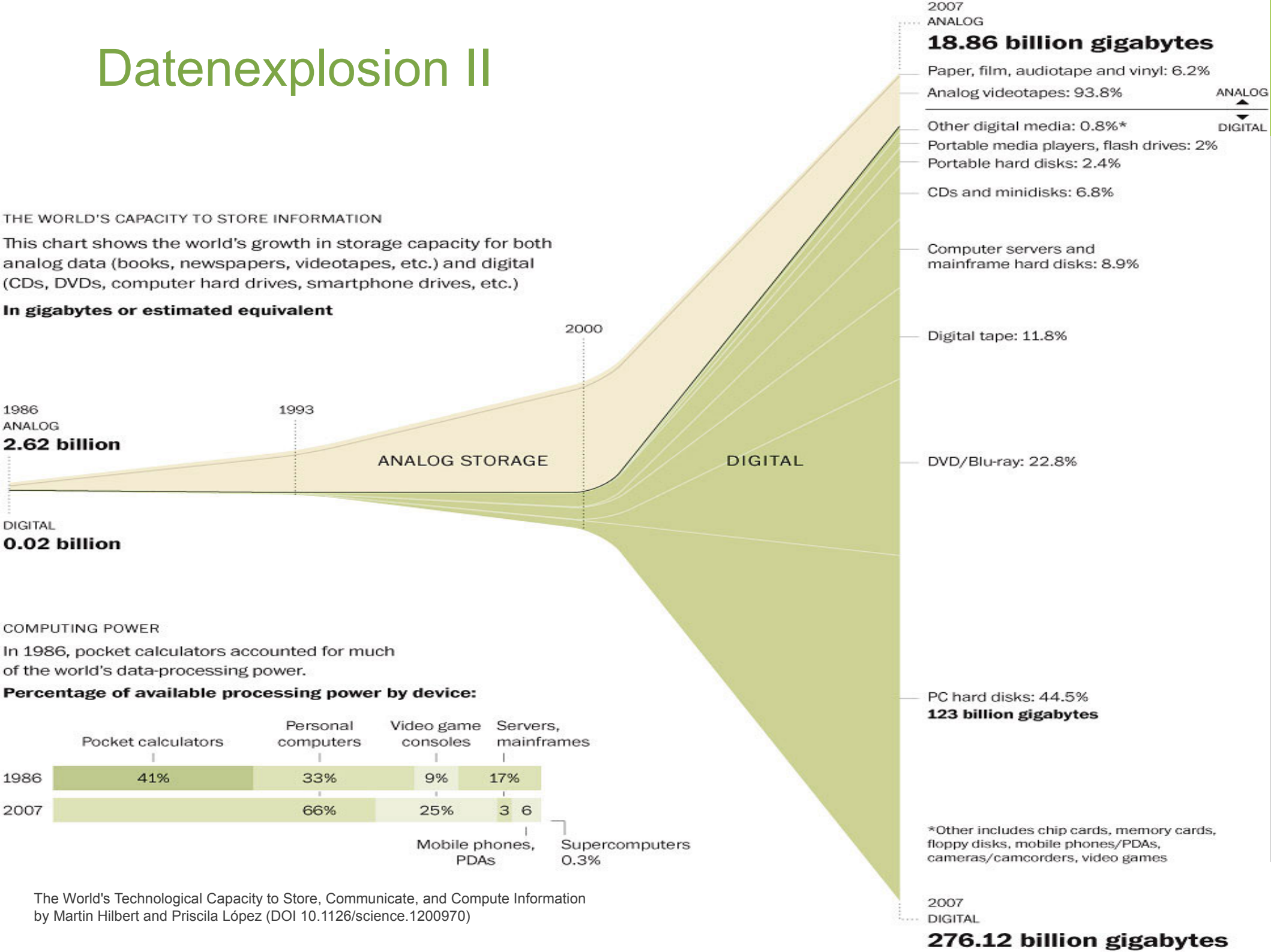
Datenexplosion I

From Bits to Zettabytes

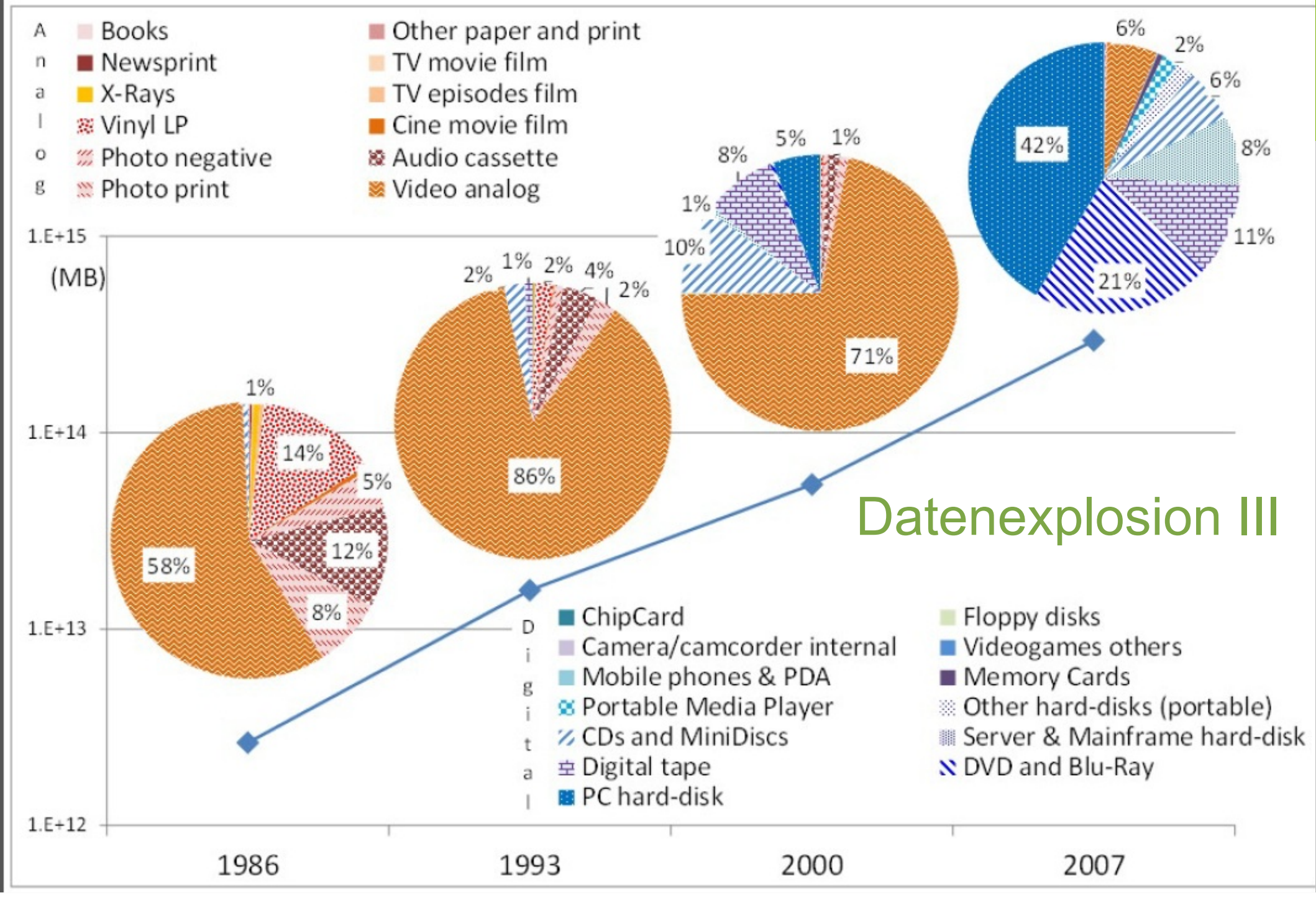
A "bit" is the smallest unit of information possible. One bit has two possible values: 1 (on) and 0 (off). A "byte" is composed of 8 bits and can represent $2^8 = 256$ values. To talk about larger amounts of data, multiples of 1000 are used: 1 Kilobyte (KB) equals 1000 bytes, 1 Megabyte (MB) equals 1000 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.



Datenexplosion II



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



Datenexplosion III

Nutzen von Modellen

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

Prozess-Modelle vs. Real-Daten: 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

Entwickelt von Motorola in frühen 1980ern.

„DMAIC“-Ansatz:

- Definiere (**D**efine) Problem und Ziel,
- Messe (**M**easure) Hauptindikatoren für Leistung und 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: Abweichung von Ziel minimieren.

Prozess “auf Six Sigma”: nur 3,4 Fehler / 1 Million Fälle, d.h. im Durchschnitt 99,9997% korrekt.

Quality level	Defective Parts per Million Opportunities (DPMO)	Percentage Passed
One Sigma	690,000 DPMO	31%
Two Sigma	308,000 DPMO	69.2%
Three Sigma	66,800 DPMO	93.32%
Four Sigma	6,210 DPMO	99.379%
Five Sigma	230 DPMO	99.977%
Six Sigma	3.4 DPMO	99.9997%

Leistungssteigerung vs. Compliance

Skandale (Enron, Tyco, Adelphia, Peregrine, WorldCom, etc.)
==>

- Zunehmendes Gewicht auf Corporate Governance, Risk und Compliance (d.h. ob Organisationen “innerhalb ihrer Grenzen” operieren), Interesse an drastischerer Auditierung.
- Neue rechtliche Regelungen:
  - z.B. 2002: Sarbanes-Oxley (SOX), 2004: Basel II

Einschränkungen von Prozessmodellen

- Kann ausführbare Modelle nutzen, um bestimmte Arbeitsweisen zu erzwingen.
- Meiste “handgemachte” Modelle unterscheiden sich von Realität und bieten höchstens idealisierte Sicht auf Prozess: “Papier-Tiger”.

Wegen:

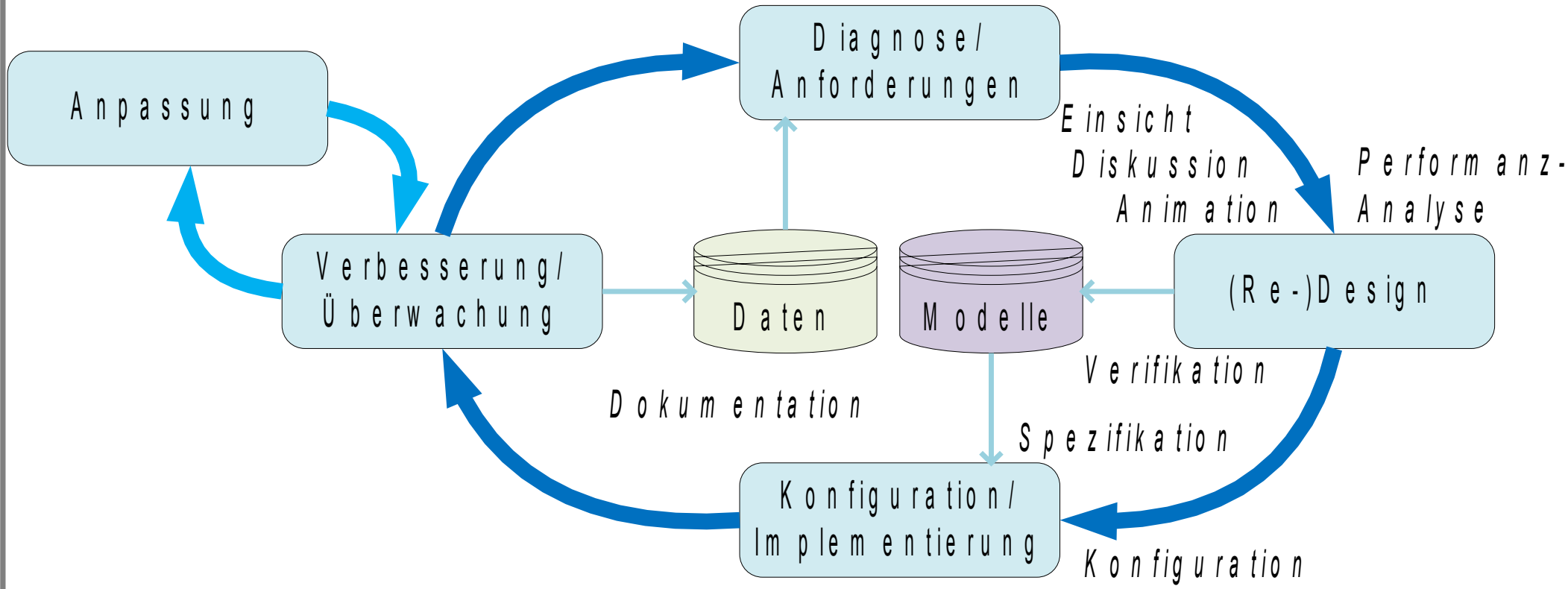
(a) Interesse an Prozessmodellen,

(b) Menge an Event-Daten und

(c) eingeschränkter Qualität von “handgemachten” Modellen:

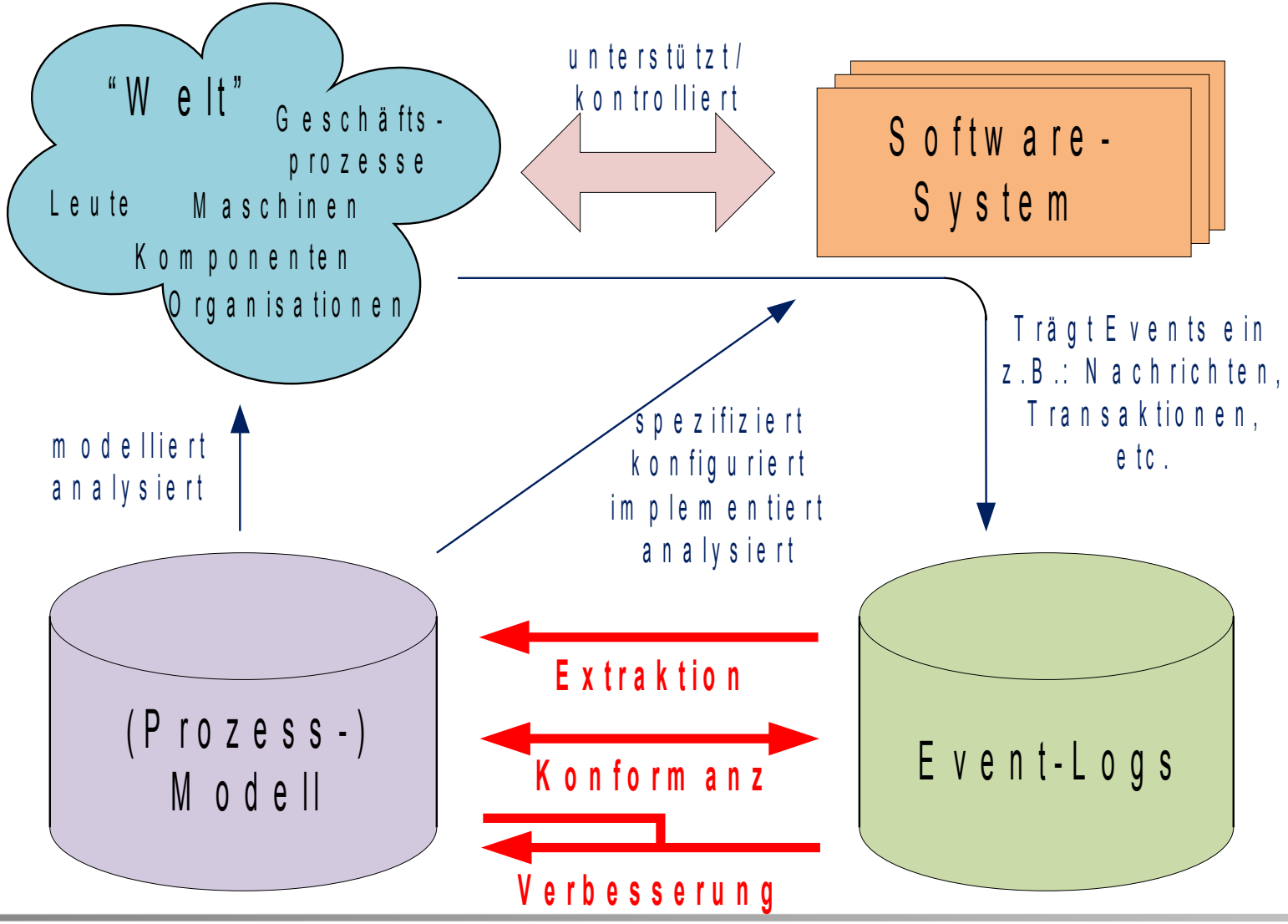
=> Event-Daten Prozessmodelle zuordnen: Process-Mining !

BPM-Lebenszyklus: klassischer Gebrauch von Prozessmodellen



Die drei Hauptarten von Process-Mining:
Extraktion, Konformanz und Verbesserung *

* discovery, conformance and enhancement



Orthogonal: Perspektiven

- Kontrollfluss-Perspektive: Reihenfolge der Aktivitäten.
- Betriebliche Perspektive: Informationen über Ressourcen (versteckt im Log), also: Welche Akteure (z.B.: Leute, Systeme, Rollen und Abteilungen) involviert, in welcher Beziehung.
- Fall-Perspektive: Eigenschaften von Fällen. Können z.B. auch über Werte der zugehörigen Daten-Elemente charakterisiert werden.
- Zeitliche Perspektive: Timing und Frequenz von Ereignissen.

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					
	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					
	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	...

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

Vereinfachter 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	…
	35654426	06-01-2011:11.18	decide	Sara	200	…
	35654427	07-01-2011:14.24	reject request	Pete	200	…
2	35654483	30-12-2010:11.32	register request	Mike	50	…
	35654485	30-12-2010:12.12	check ticket	Mike	100	…
	35654487	30-12-2010:14.16	examine casually	Pete	400	…
	35654488	05-01-2011:11.22	decide	Sara	200	…
	35654489	08-01-2011:12.05	pay compensation	Ellen	200	…
3	35654521	30-12-2010:14.32	register request	Pete	50	…
	35654522	30-12-2010:15.06	examine casually	Mike	400	…
	35654524	30-12-2010:16.34	check ticket	Ellen	100	…
	35654525	06-01-2011:09.18	decide	Sara	200	…
	35654526	06-01-2011:12.18	reinitiate request	Sara	200	…
	35654527	06-01-2011:13.06	examine thoroughly	Sean	400	…
	35654530	08-01-2011:11.43	check ticket	Pete	100	…
	35654531	09-01-2011:09.55	decide	Sara	200	…
	35654533	15-01-2011:10.45	pay compensation	Ellen	200	…
4	35654641	06-01-2011:15.02	register request	Pete	50	…
	35654643	07-01-2011:12.06	check ticket	Mike	100	…
	35654644	08-01-2011:14.43	examine thoroughly	Sean	400	…
	35654645	09-01-2011:12.02	decide	Sara	200	…
	35654647	12-01-2011:15.44	reject request	Ellen	200	…
5	35654711	06-01-2011:09.02	register request	Ellen	50	…
	35654712	07-01-2011:10.16	examine casually	Mike	400	…
	35654714	08-01-2011:11.22	check ticket	Pete	100	…
	35654715	10-01-2011:13.28	decide	Sara	200	…
	35654716	11-01-2011:16.18	reinitiate request	Sara	200	…
	35654718	14-01-2011:14.33	check ticket	Ellen	100	…
	35654719	16-01-2011:15.50	examine casually	Mike	400	…
	35654720	19-01-2011:11.18	decide	Sara	200	…
	35654721	20-01-2011:12.48	reinitiate request	Sara	200	…
	35654722	21-01-2011:09.06	examine casually	Sue	400	…
	35654724	21-01-2011:11.34	check ticket	Pete	100	…
	35654725	23-01-2011:13.12	decide	Sara	200	…
	35654726	24-01-2011:14.56	reject request	Mike	200	…
6	35654871	06-01-2011:15.02	register request	Mike	50	…
	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	…
…	…	…	…	…	…	…

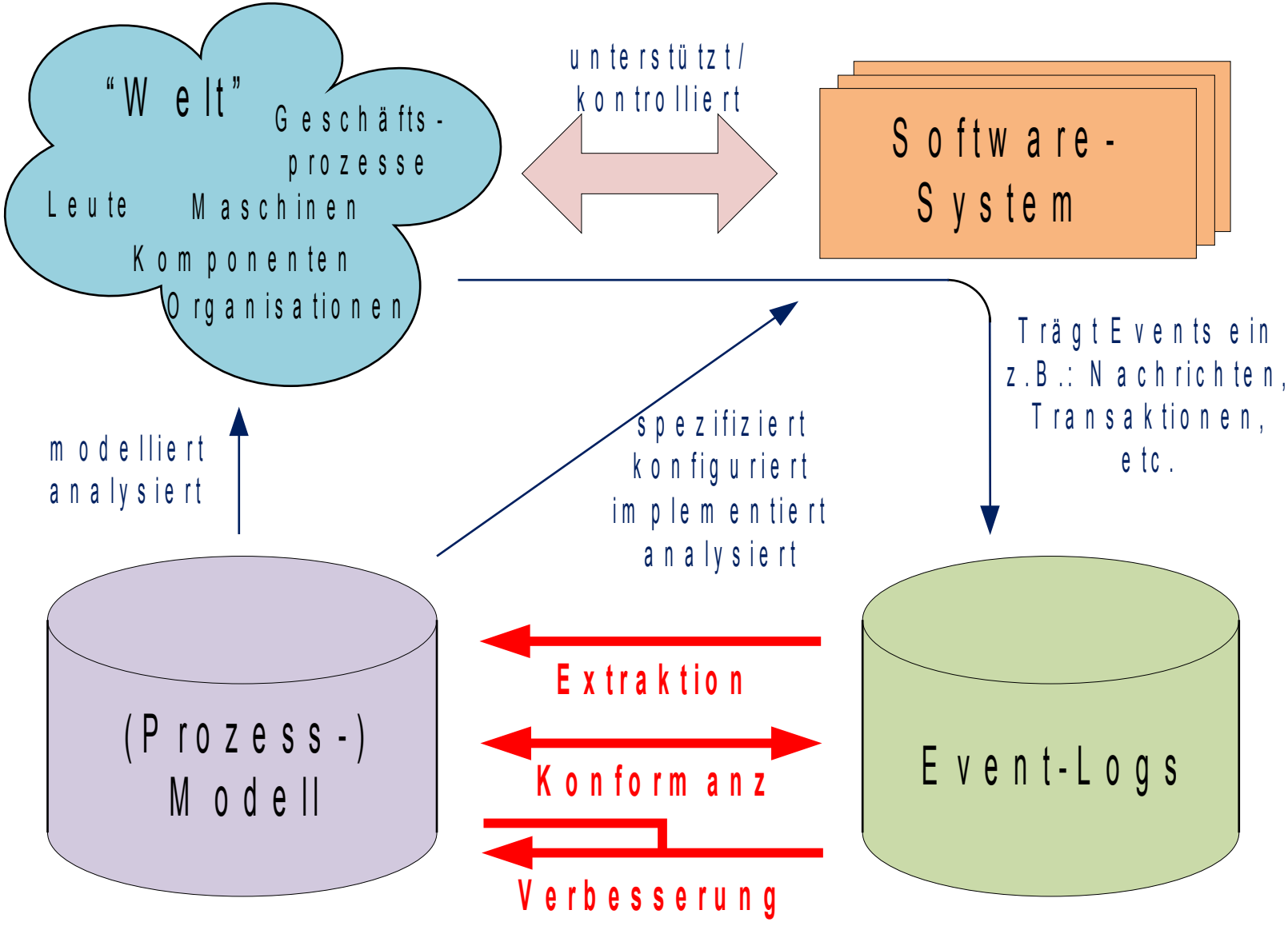
case id	trace
1	$\langle a, b, d, e, h \rangle$
2	$\langle a, d, c, e, g \rangle$
3	$\langle a, c, d, e, f, b, d, e, g \rangle$
4	$\langle a, d, b, e, h \rangle$
5	$\langle a, c, d, e, f, d, c, e, f, c, d, e, h \rangle$
6	$\langle a, c, d, e, g \rangle$
…	…

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)

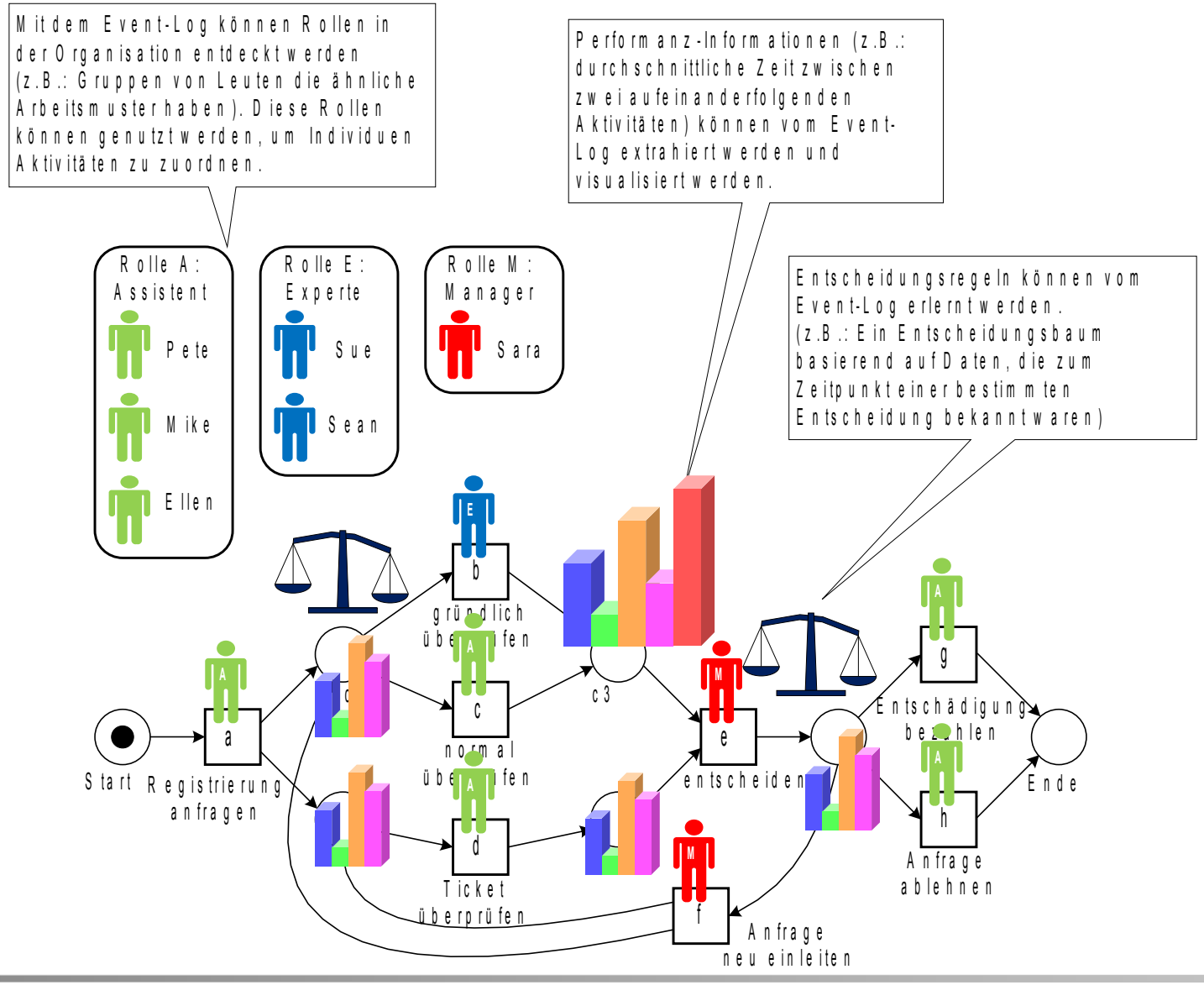
Prozess-Discovery: BPMN-Modell

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	$\langle a, c, d, e, g \rangle$
...	...

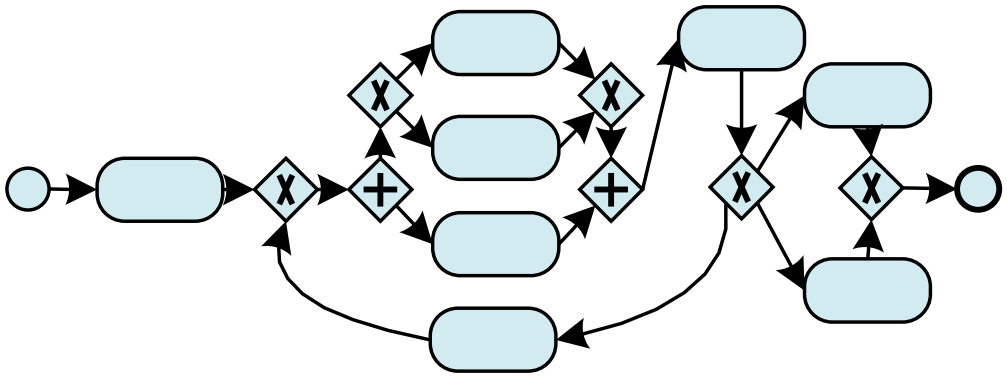
Über Extraktion hinaus: Konformanz und Verbesserung



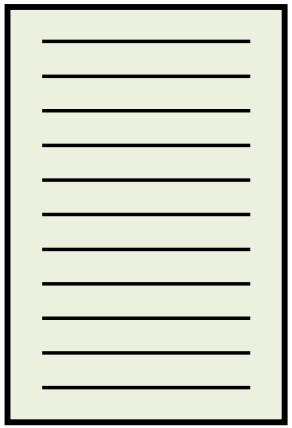
Erweiterung



Play-Out

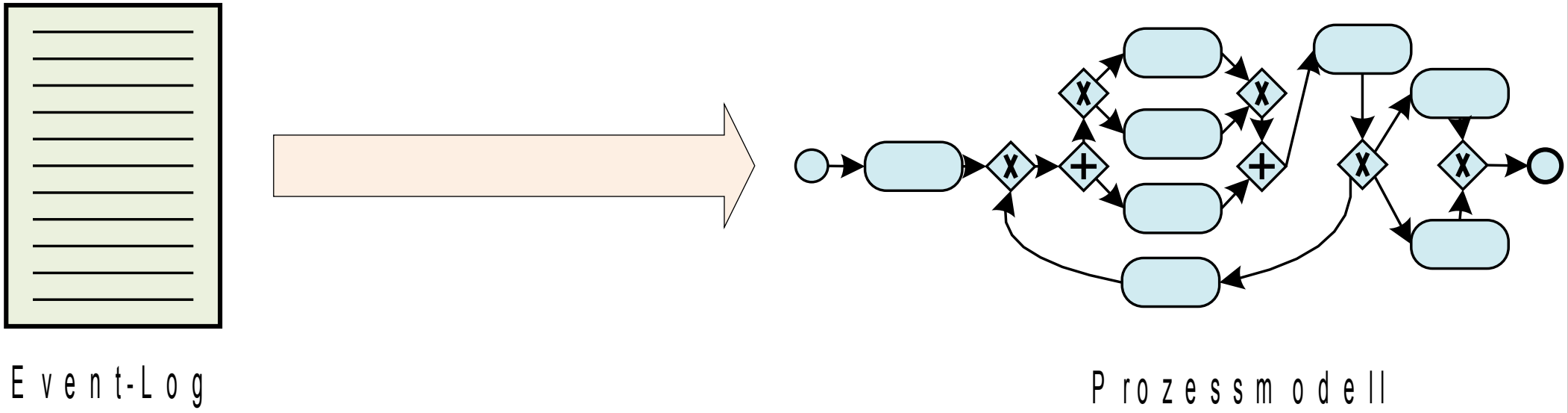


Prozessmodell

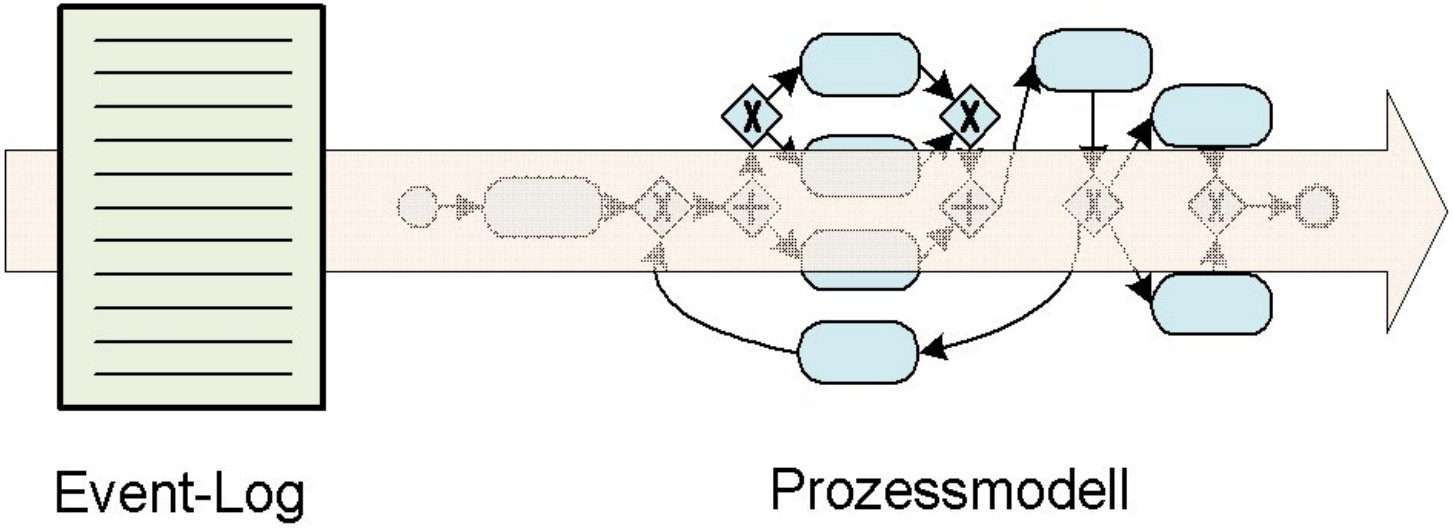


Event-Log

Play-In



Replay



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

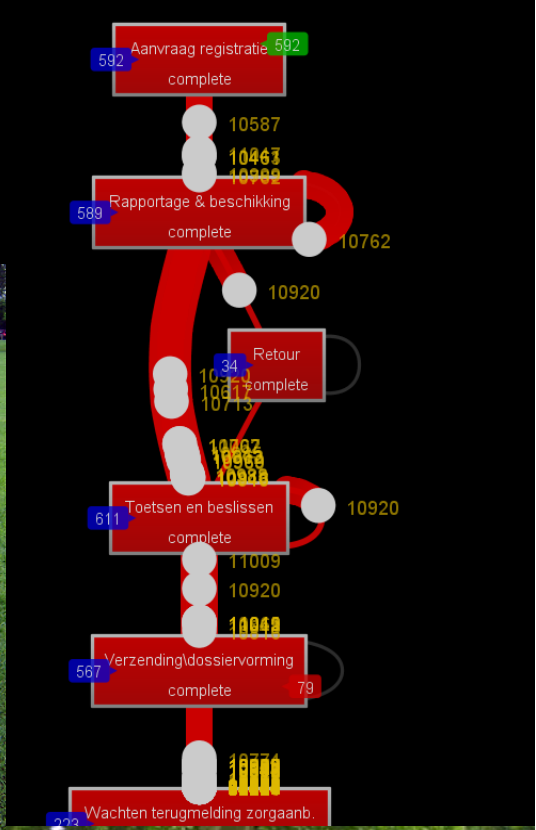
Replay

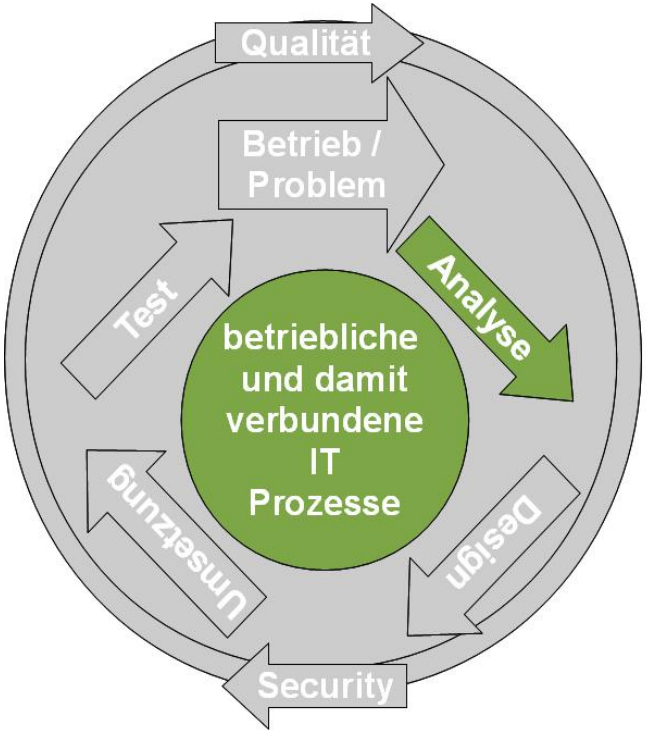
Wichtig: **Modelle** mit **echten Ereignissen** verknüpfen.

Mögliche Anwendungen:

- **Erfüllung** prüfen
- Modelle **reparieren**
- Modell um Informationen zu **Frequenz** und **Zeit** erweitern
- **Voraussagende** Modelle konstruieren
- **Betriebsunterstützung** (Vorhersagen, Vorschläge, etc.)

Wunschpfade in Prozess-Modellen



Zusammenfassung
2.0: Einführung Business Process Mining

- Geschäftsprozessmodellierung
- Process Mining
  - Einführung: Process Mining
- Modellbasierte Softwareentwicklung
- Modellbasierte Entwicklung sicherer Software