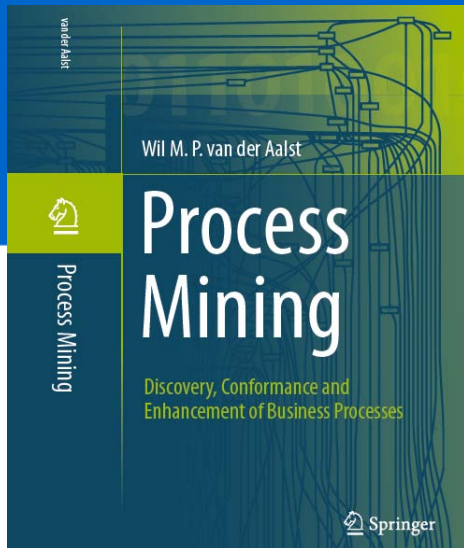


# Chapter 1

# Introduction

prof.dr.ir. Wil van der Aalst  
[www.processmining.org](http://www.processmining.org)



**TU** / **e** Technische Universiteit  
**Eindhoven**  
University of Technology

Where innovation starts

# Overview

Chapter 1  
Introduction

---

*Part I: Preliminaries*

Chapter 2  
Process Modeling and  
Analysis

Chapter 3  
Data Mining

---

*Part II: From Event Logs to Process Models*

Chapter 4  
Getting the Data

Chapter 5  
Process Discovery: An  
Introduction

Chapter 6  
Advanced Process  
Discovery Techniques

---

*Part III: Beyond Process Discovery*

Chapter 7  
Conformance  
Checking

Chapter 8  
Mining Additional  
Perspectives

Chapter 9  
Operational Support

---

*Part IV: Putting Process Mining to Work*

Chapter 10  
Tool Support

Chapter 11  
Analyzing “Lasagna  
Processes”

Chapter 12  
Analyzing “Spaghetti  
Processes”

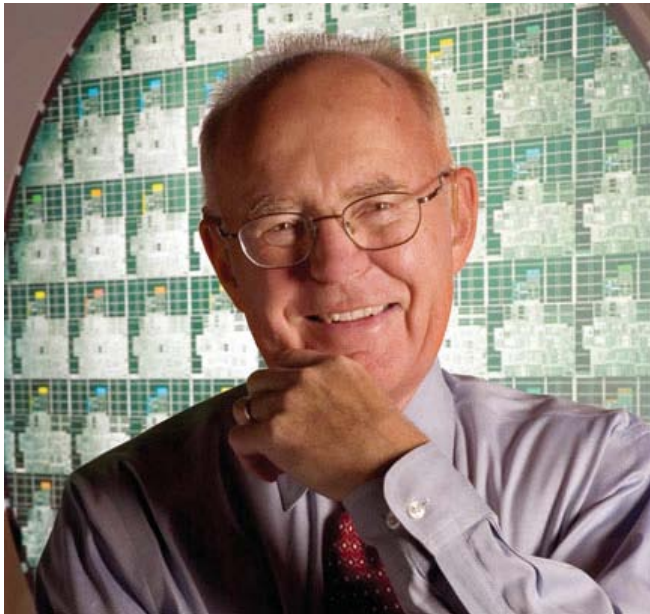
---

*Part V: Reflection*

Chapter 13  
Cartography and  
Navigation

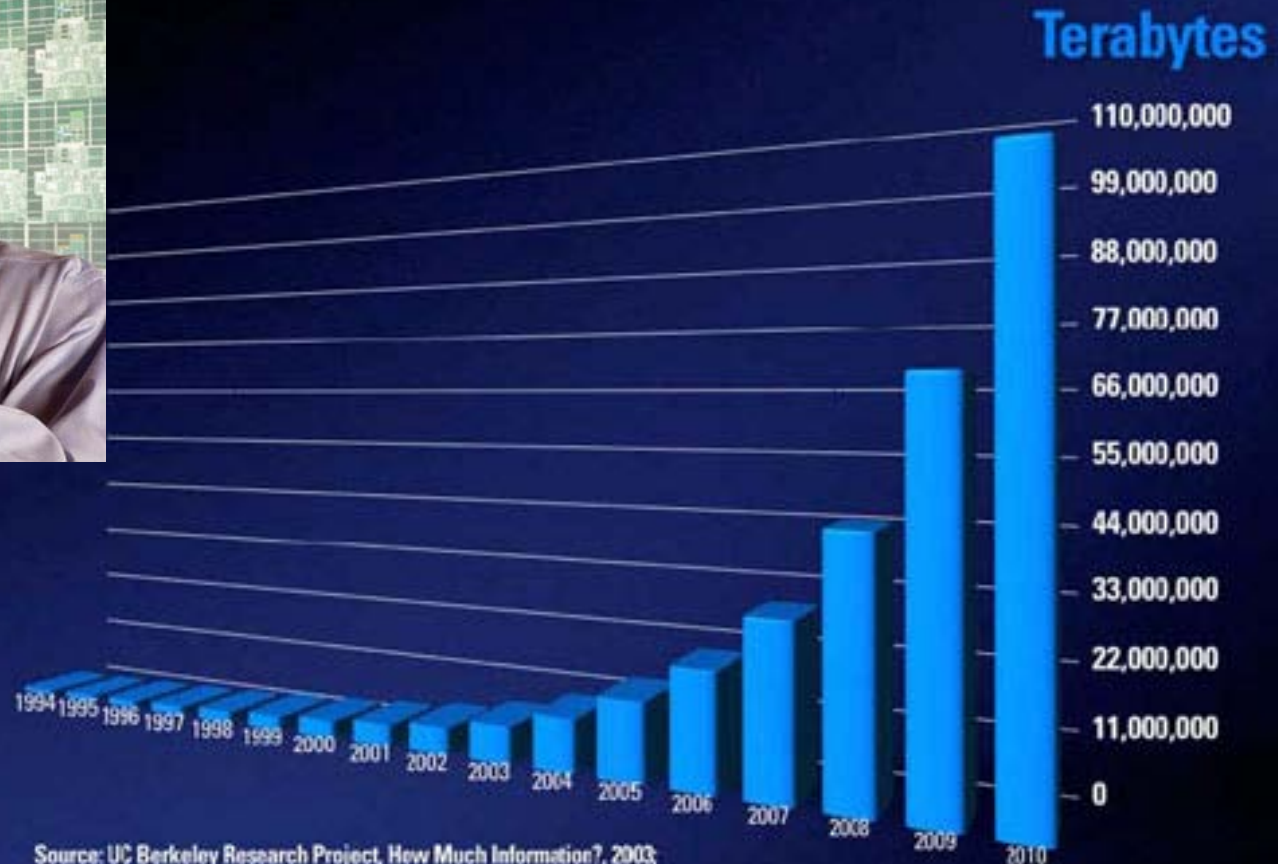
Chapter 14  
Epilogue

# Data explosion



## 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  $10^{21} = 1,000,000,000,000,000,000$  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  $2^{20} = 1,048,576$  bytes, and 1 Zettabyte is  $2^{70} \approx 1.18 \times 10^{21}$  bytes.



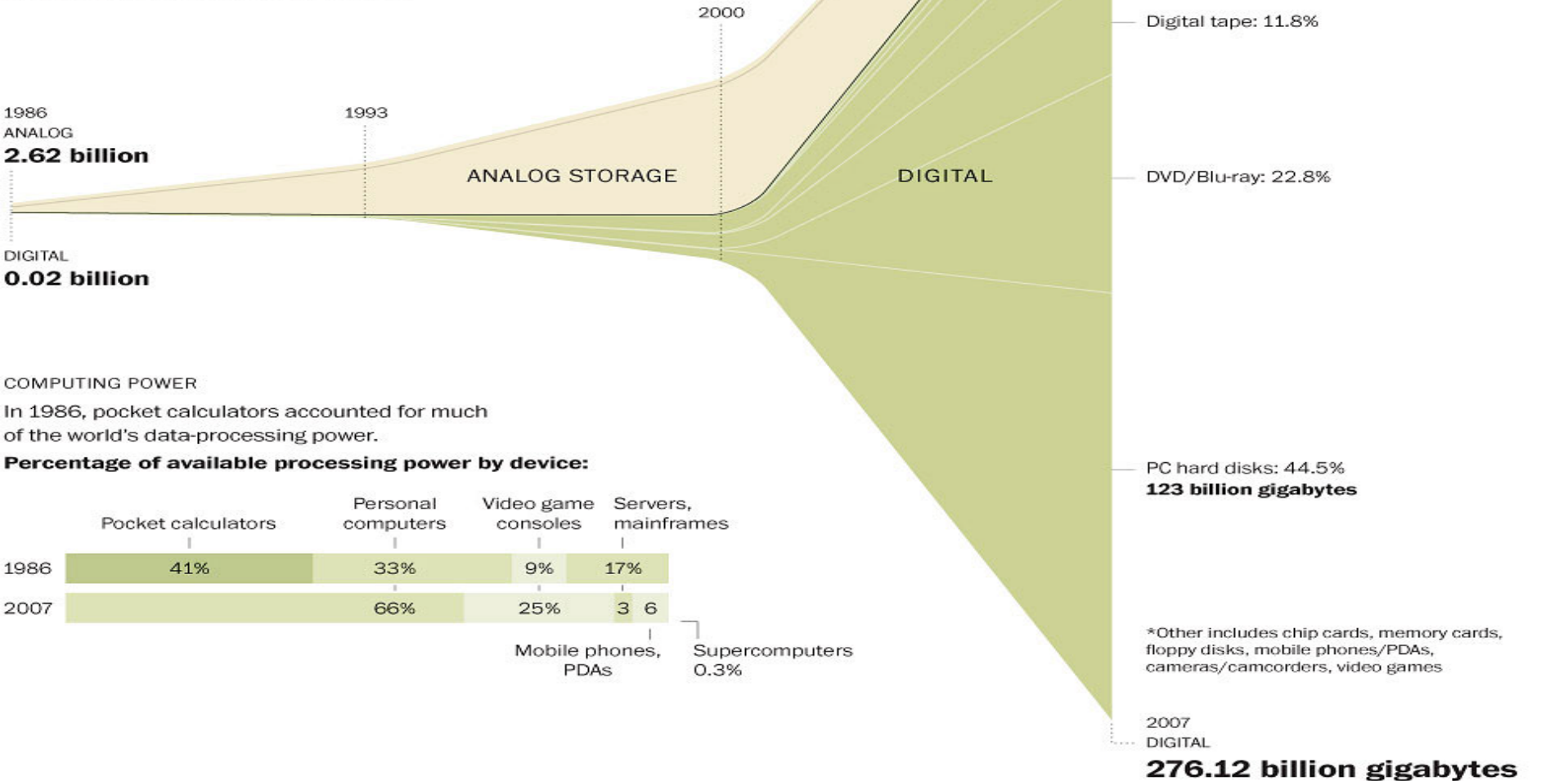
Source: UC Berkeley Research Project, How Much Information?, 2003;  
IDC, Disk Storage System Quarterly Tracker (as of 2006)

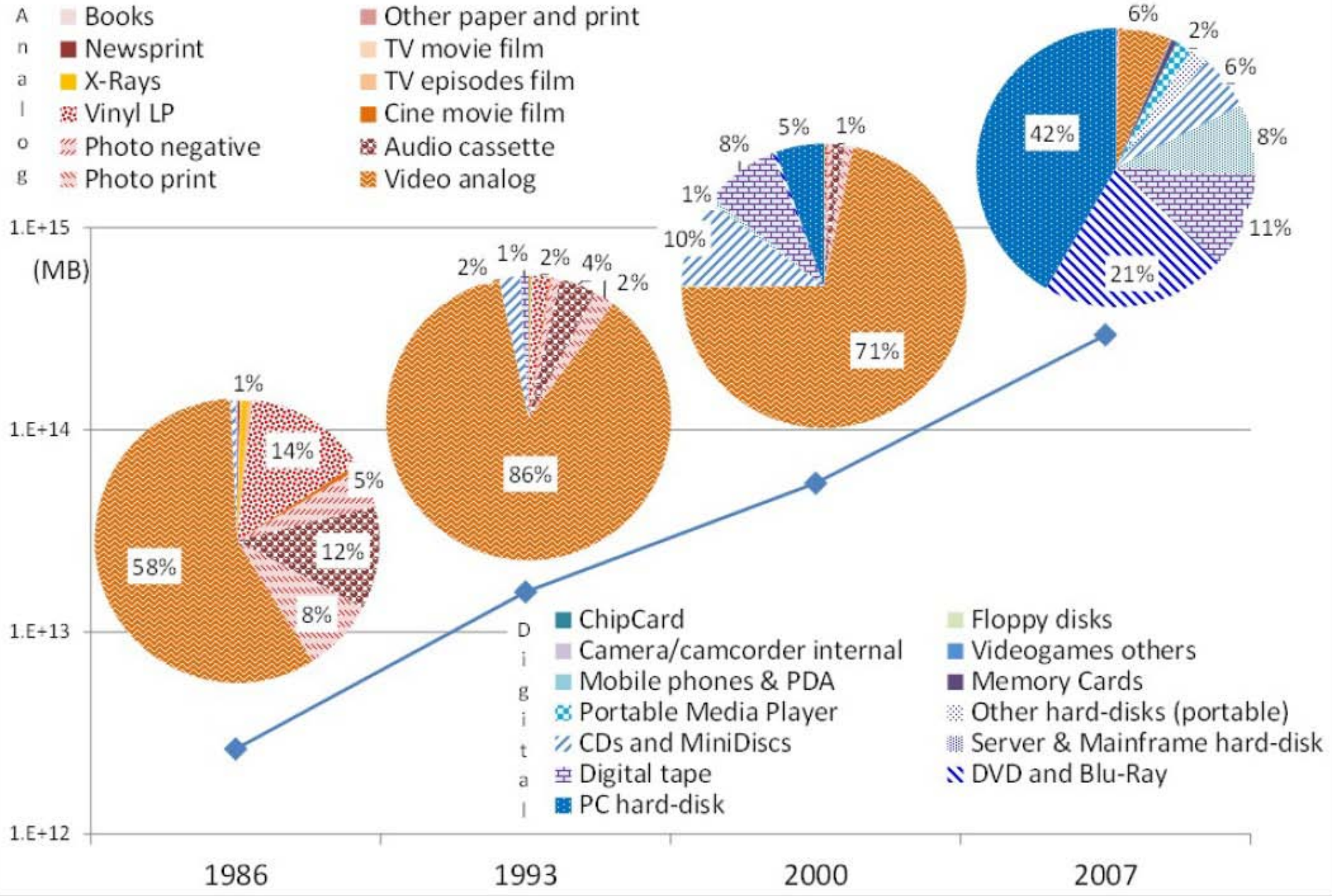
**The World's Technological Capacity to Store, Communicate, and Compute Information by Martin Hilbert and Priscila López (DOI 10.1126/science.1200970)**

**THE WORLD'S CAPACITY TO STORE INFORMATION**

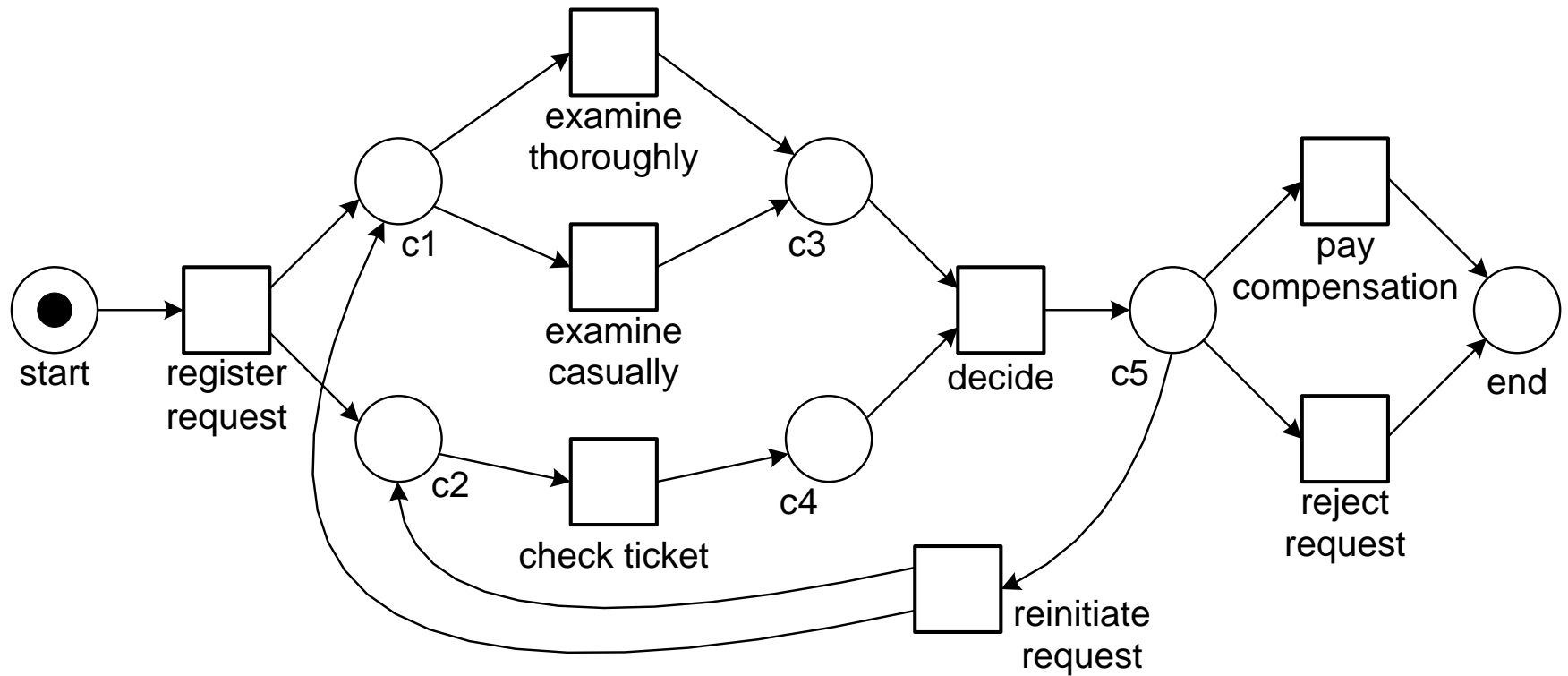
This chart shows the world's growth in storage capacity for both analog data (books, newspapers, videotapes, etc.) and digital (CDs, DVDs, computer hard drives, smartphone drives, etc.)

**In gigabytes or estimated equivalent**

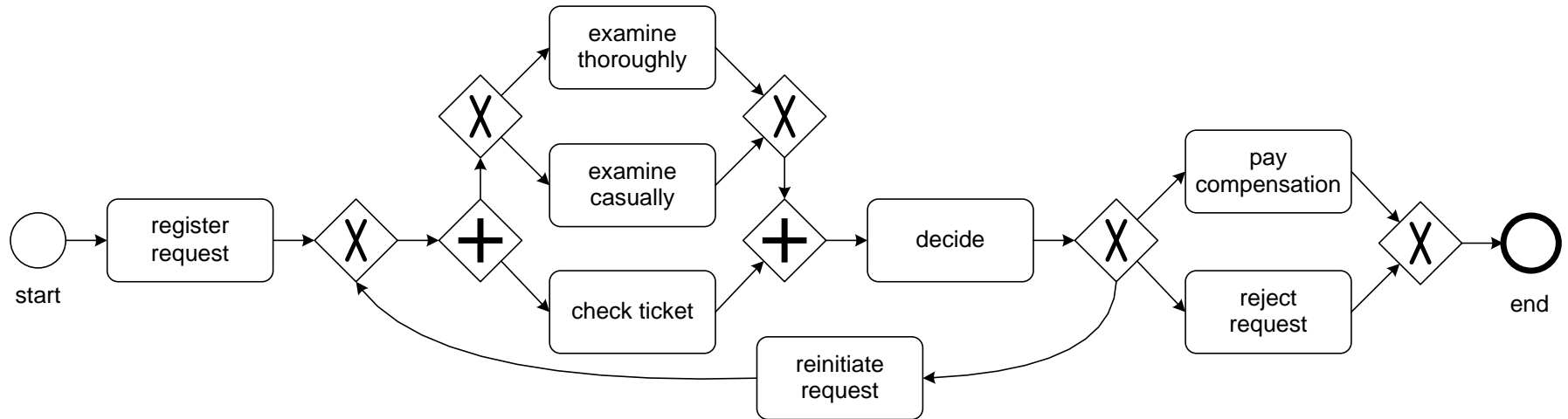




# Example process model



# Same process in terms of BPMN rather than Petri nets



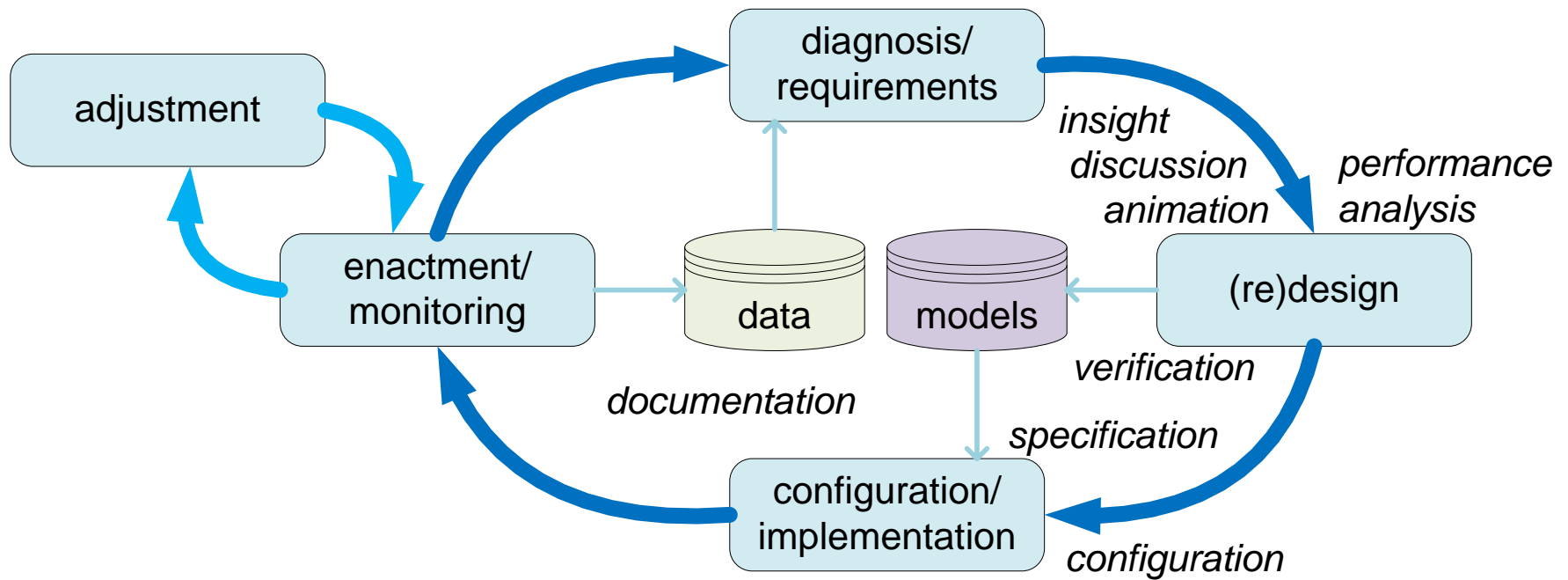
# What are process models used for?

- **insight**: while making a model, the modeler is triggered to view the process from various angles;
- **discussion**: the stakeholders use models to structure discussions;
- **documentation**: processes are documented for instructing people or certification purposes (cf. ISO 9000 quality management);
- **verification**: process models are analyzed to find errors in systems or procedures (e.g., potential deadlocks);
- **performance analysis**: techniques like simulation can be used to understand the factors influencing response times, service levels, etc.;
- **animation**: models enable end users to “play out” different scenarios and thus provide feedback to the designer;
- **specification**: models can be used to describe a PAIS before it is implemented and can hence serve as a “contract” between the developer and the end user/management; and
- **configuration**: models can be used to configure a system.

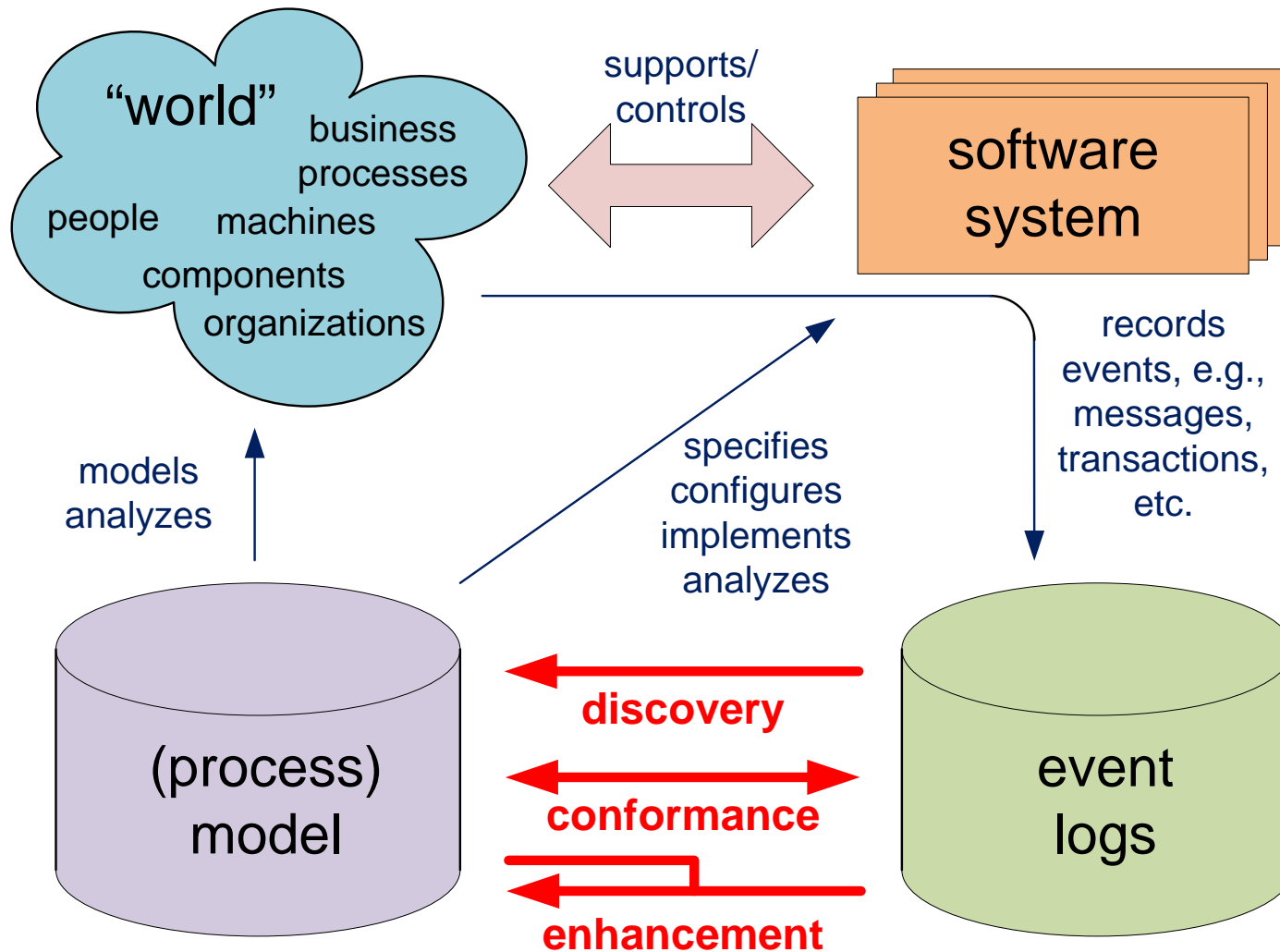
# Limitations

- Executable models may be used to force people to work in a particular manner.
- However, most models are **not well-aligned** with reality.
- Most hand-made models are **disconnected from reality** and provide only an idealized view on the processes at hand: “paper tigers”.
- Given (a) the interest in process models, (b) the abundance of event data, and (c) the limited quality of hand-made models, it seems worthwhile to relate event data to process models: **process mining!**

# BPM life-cycle showing the classical uses of process models



# The three main types of process mining: discovery, conformance, and enhancement



# Orthogonal: Perspectives

- The **control-flow perspective** focuses on the control-flow, i.e., the ordering of activities.
- The **organizational perspective** focuses on information about resources hidden in the log, i.e., which actors (e.g., people, systems, roles, and departments) are involved and how are they related.
- The **case perspective** focuses on properties of cases, e.g., cases can also be characterized by the values of the corresponding data elements.
- The **time perspective** is concerned with the timing and frequency of events.

# Starting point: 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	Pete	400	...
	35654524	30-12-2010:16.34	check ticket	Mike	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	Sue	400	...
	35654530	08-01-2011:11.43	check ticket	Mike	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	Sue	400	...
	35654645	09-01-2011:12.02	decide	Sara	200	...
35654647	12-01-2011:15.44	reject request	Pete	200	...	
5	35654711	06-01-2011:09.02	register request	Pete	50	...
	35654712	07-01-2011:10.16	examine casually	Pete	400	...
	35654714	08-01-2011:11.22	check ticket	Mike	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	Mike	100	...
	35654719	16-01-2011:15.50	examine casually	Pete	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	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	...

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

# Simplified event log

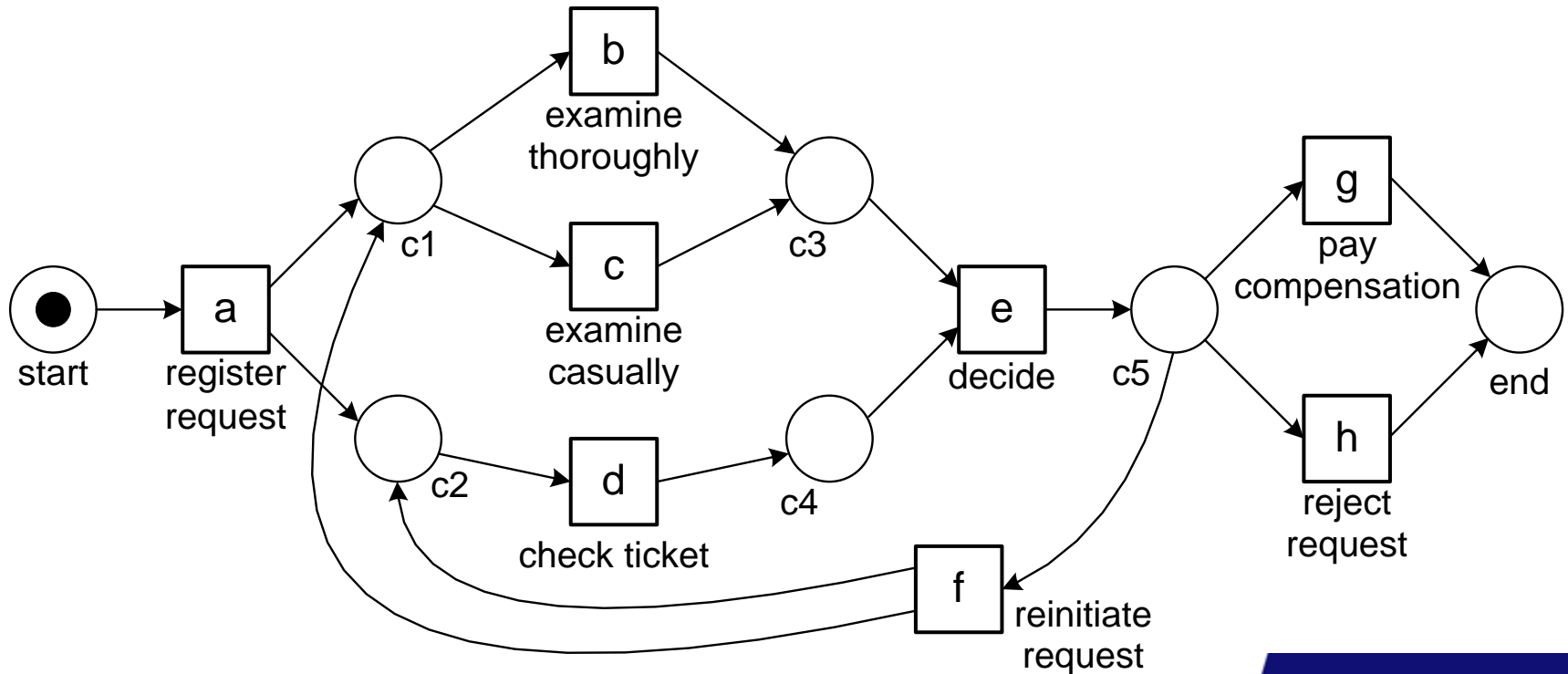
case id	event id	properties		
		timestamp	activity	resource
1	35654423	30-12-2010:11.02	register request	Pete
	35654424	31-12-2010:10.06	examine thoroughly	Sue
	35654425	05-01-2011:15.12	check ticket	Mike
	35654426	06-01-2011:11.18	decide	Sara
	35654427	07-01-2011:14.24	reject request	Pete
2	35654483	30-12-2010:11.32	register request	Mike
	35654485	30-12-2010:12.12	check ticket	Mike
	35654487	30-12-2010:14.16	examine casually	Pete
	35654488	05-01-2011:11.22	decide	Sara
	35654489	08-01-2011:12.05	pay compensation	Ellen
3	35654521	30-12-2010:14.32	register request	Pete
	35654522	30-12-2010:15.06	examine casually	Mike
	35654524	30-12-2010:16.34	check ticket	Ellen
	35654525	06-01-2011:09.18	decide	Sara
	35654526	06-01-2011:12.18	reinitiate request	Sara
	35654527	06-01-2011:13.06	examine thoroughly	Sean
	35654530	08-01-2011:11.43	check ticket	Pete
	35654531	09-01-2011:09.55	decide	Sara
	35654533	15-01-2011:10.45	pay compensation	Ellen
4	35654641	06-01-2011:15.02	register request	Pete
	35654643	07-01-2011:12.06	check ticket	Mike
	35654644	08-01-2011:14.43	examine thoroughly	Sean
	35654645	09-01-2011:12.02	decide	Sara
	35654647	12-01-2011:15.44	reject request	Ellen
5	35654711	06-01-2011:09.02	register request	Ellen
	35654712	07-01-2011:10.16	examine casually	Mike
	35654714	08-01-2011:11.22	check ticket	Pete
	35654715	10-01-2011:13.28	decide	Sara
	35654716	11-01-2011:16.18	reinitiate request	Sara
	35654718	14-01-2011:14.33	check ticket	Ellen
	35654719	16-01-2011:15.50	examine casually	Mike
	35654720	19-01-2011:11.18	decide	Sara
	35654721	20-01-2011:12.48	reinitiate request	Sara
	35654722	21-01-2011:09.06	examine casually	Sue
	35654724	21-01-2011:11.34	check ticket	Pete
	35654725	23-01-2011:13.12	decide	Sara
35654726	24-01-2011:14.56	reject request	Mike	
6	35654871	06-01-2011:15.02	register request	Mike
	35654873	06-01-2011:16.06	examine casually	Ellen
	35654874	07-01-2011:16.22	check ticket	Mike
	35654875	07-01-2011:16.52	decide	Sara
	35654877	16-01-2011:11.47	pay compensation	Mike
...	...	...	...	...

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 = register request,**  
**b = examine thoroughly,**  
**c = examine casually,**  
**d = check ticket,**  
**e = decide,**  
**f = reinitiate request,**  
**g = pay compensation,**  
**and h = reject request**

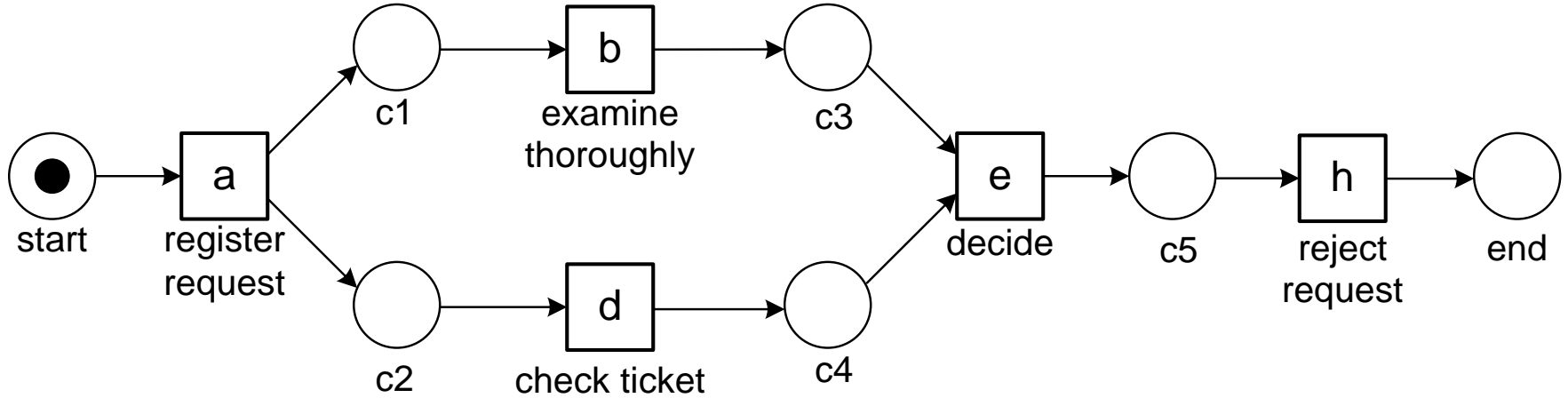
# Process discovery

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

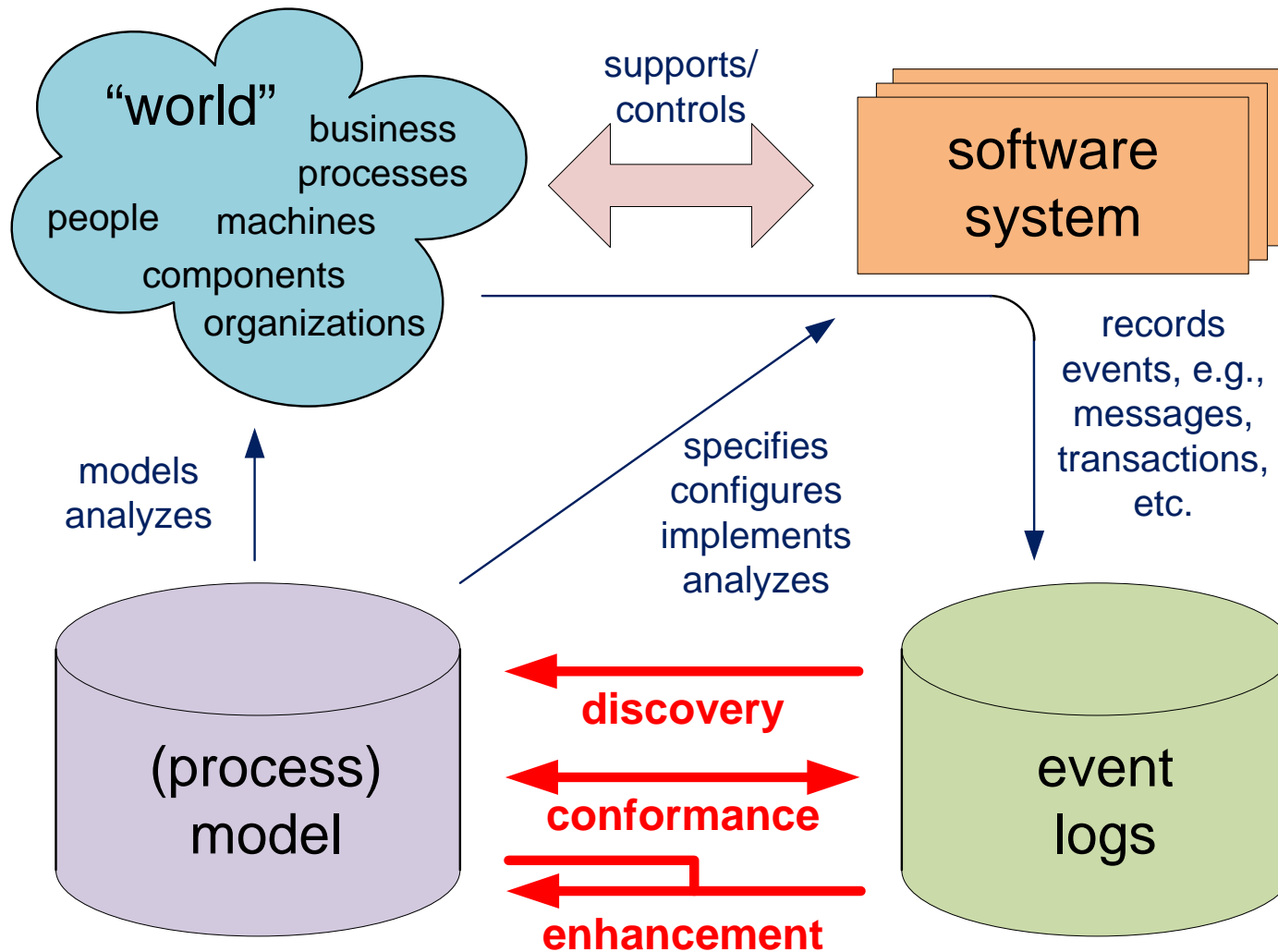


# Another example

$\{\langle a, b, d, e, h \rangle, \langle a, d, b, e, h \rangle\}$

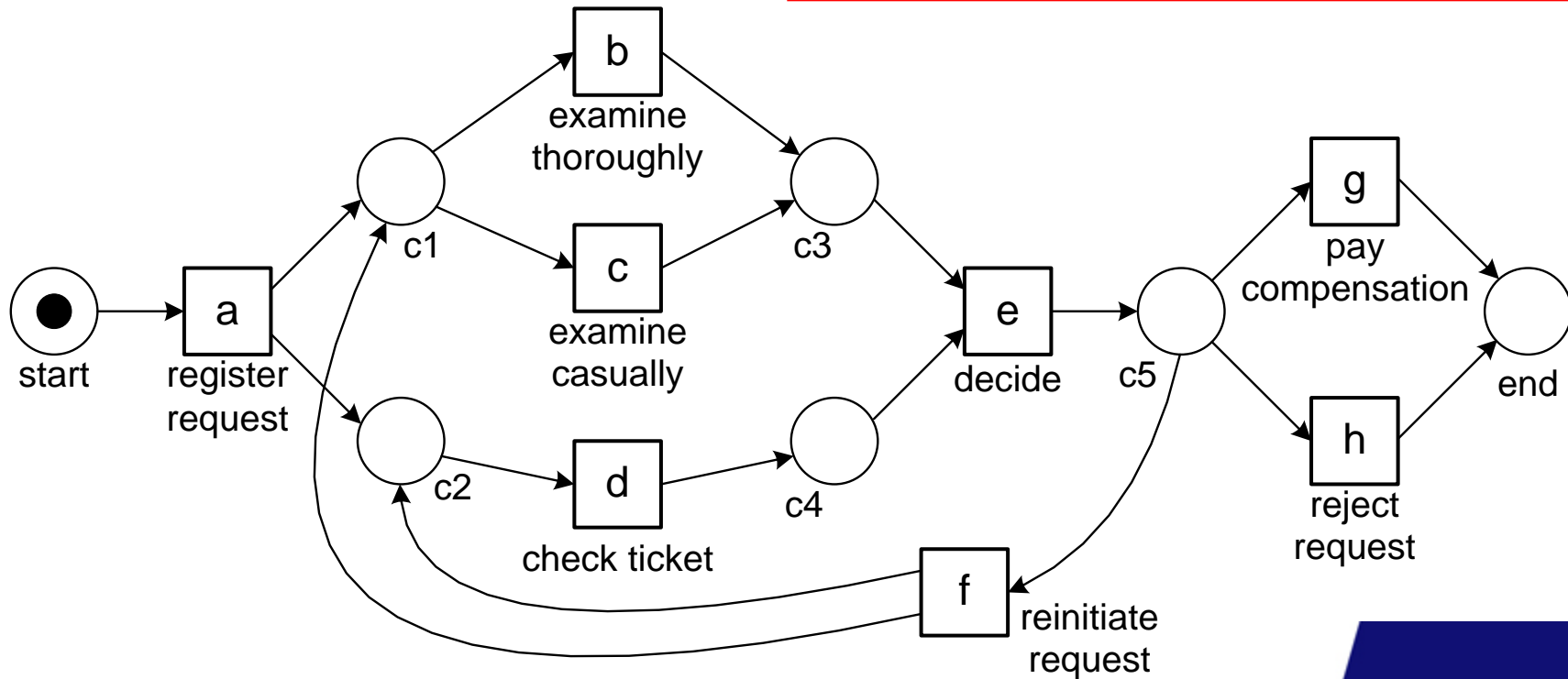


# Beyond discovery: conformance and enhancement



# Another event log

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	$\langle a, c, d, e, g \rangle$
7	$\langle a, b, e, g \rangle$
8	$\langle a, b, d, e \rangle$
9	$\langle a, d, c, e, f, d, c, e, f, b, d, e, h \rangle$
10	$\langle a, c, d, e, f, b, d, g \rangle$

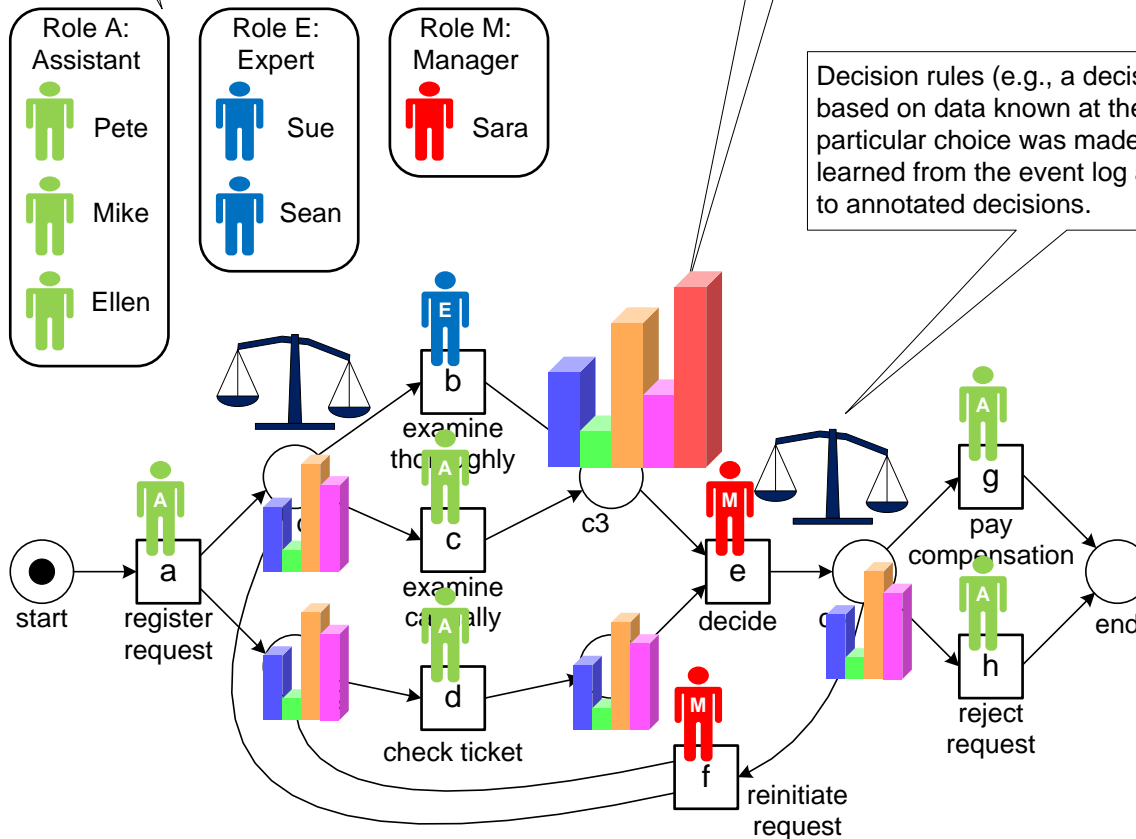


# Extension

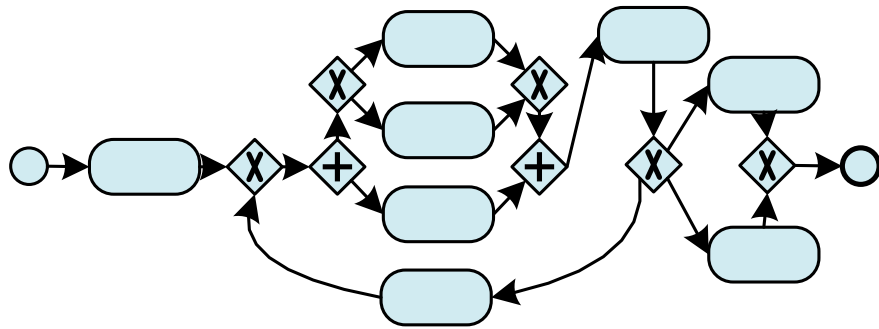
The event log can be used to discover roles in the organization (e.g., groups of people with similar work patterns). These roles can be used to relate individuals and activities.

Performance information (e.g., the average time between two subsequent activities) can be extracted from the event log and visualized on top of the model.

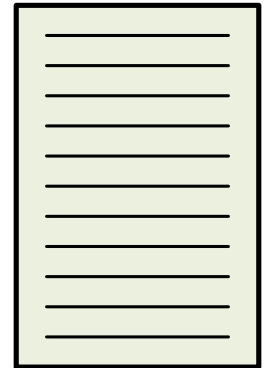
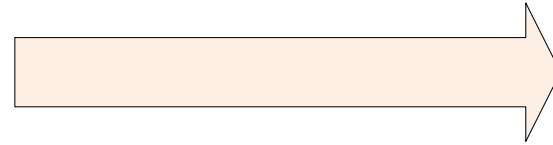
Decision rules (e.g., a decision tree based on data known at the time a particular choice was made) can be learned from the event log and used to annotated decisions.



# Play-Out

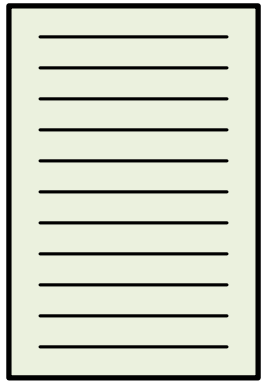


process model

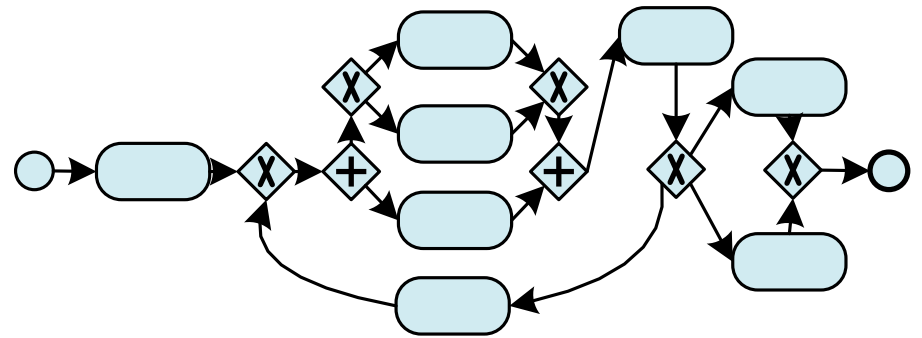
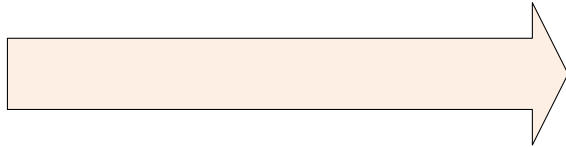


event log

# Play-In

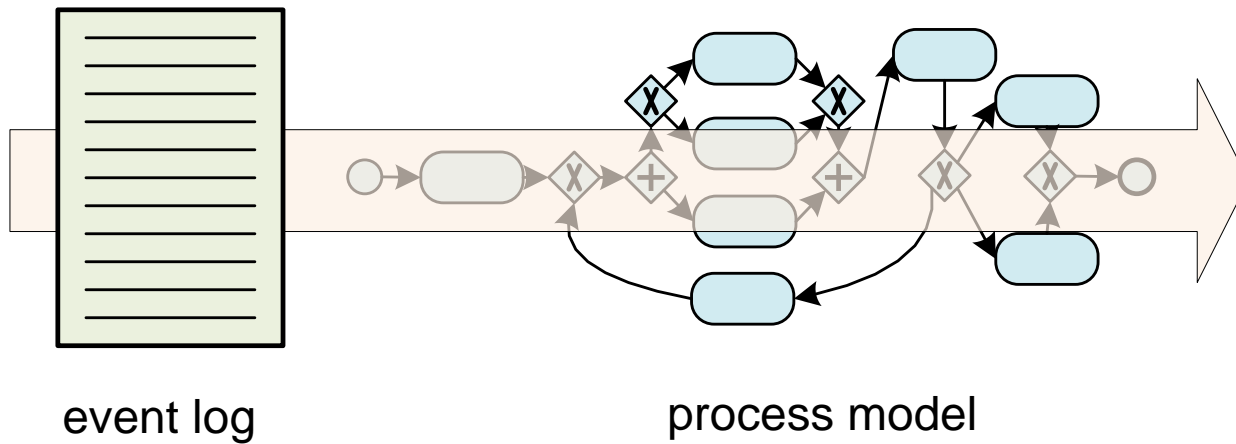


event log



process model

# Replay

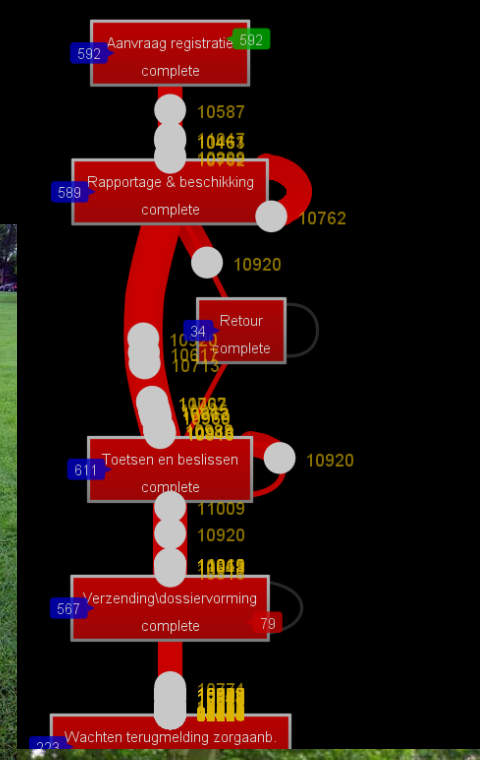


- extended model showing times, frequencies, etc.
- diagnostics
- predictions
- recommendations

# Replay

- **Connecting models to real events is crucial!**
- **Possible uses:**
  - **Conformance checking**
  - **Repairing models**
  - **Extending the model with frequencies and temporal information**
  - **Constructing predictive models**
  - **Operational support (prediction, recommendation, etc.)**

# Desire lines in process models



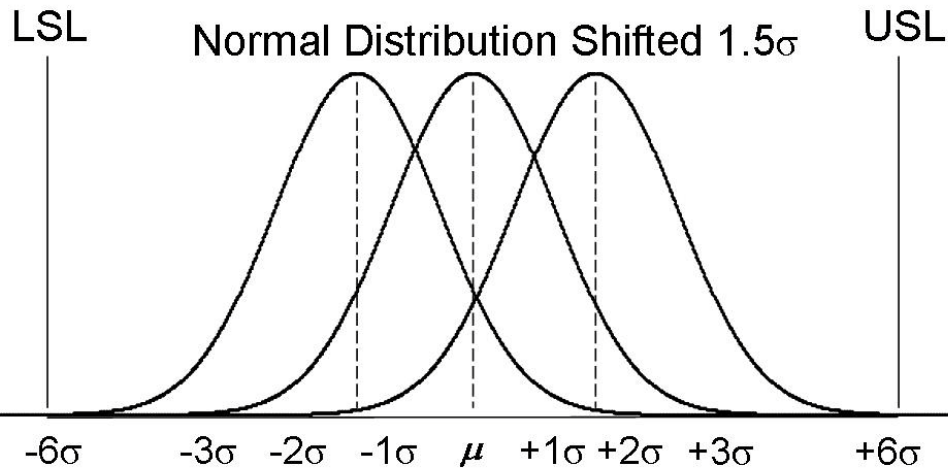
# Trends and terms

- **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**

# Six Sigma

- **Six Sigma** was originally developed by Motorola in the early 1980s.
- **DMAIC** approach:
  - **Define** the problem and set targets,
  - **Measure** key performance indicators and collect data,
  - **Analyze** the data to investigate and verify cause-and-effect relationships,
  - **Improve** the current process based on this analysis,
  - **Control** the process to minimize deviations from the target.

# $[\mu - 6\sigma, \mu + 6\sigma]$ with a $1.5\sigma$ shift



**A process that “runs at Six Sigma” has only 3.4 defective cases per million cases, i.e., on average 99.9997% of the cases is handled properly.**

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%

# Performance improvement versus compliance

- Organizations are also putting more emphasis on **corporate governance, risk, and compliance**.
- Scandals (Enron, Tyco, Adelphia, Peregrine, WorldCom, etc.) have fueled interest in more rigorous auditing practices.
- New legislation such as the **Sarbanes-Oxley Act (SOX)** of 2002 and the **Basel II Accord** of 2004 emerged as a result.
- Importance of verifying whether organizations operate “within their boundaries” is increasing.

# Outlook

Chapter 1  
Introduction

---

*Part I: Preliminaries*

Chapter 2  
Process Modeling and  
Analysis

Chapter 3  
Data Mining

---

*Part II: From Event Logs to Process Models*

Chapter 4  
Getting the Data

Chapter 5  
Process Discovery: An  
Introduction

Chapter 6  
Advanced Process  
Discovery Techniques

---

*Part III: Beyond Process Discovery*

Chapter 7  
Conformance  
Checking

Chapter 8  
Mining Additional  
Perspectives

Chapter 9  
Operational Support

---

*Part IV: Putting Process Mining to Work*

Chapter 10  
Tool Support

Chapter 11  
Analyzing “Lasagna  
Processes”

Chapter 12  
Analyzing “Spaghetti  
Processes”

---

*Part V: Reflection*

Chapter 13  
Cartography and  
Navigation

Chapter 14  
Epilogue