Management und Analyse von Provenancedaten

Ulf Leser

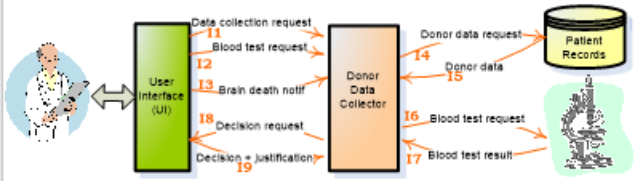
Luc Moreau, Open Provenance Model Tutorial

Provenance Use Cases

Open Provenance Model

- Which doctor was involved in a decision?
- Why an organ was rejected for transplant?
- Was an organ allocated according to rules?

- Was the data used in a manner compatible with the purpose it was captured for?
- Was the latest data used in the computation?
- Was the data deleted after its use?



Organ Transplant Management
(Vazquez Salceda, Willmott 05-07)

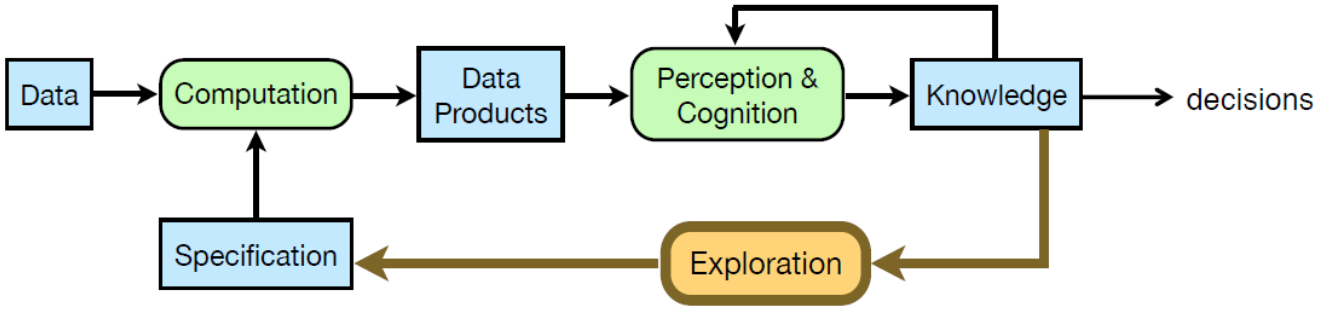
Auditing of private data processing
(Rocio Aldeco Perez 08)

For an extensive catalogue of provenance use cases, see W3C incubator

J. Freire, 2021

Data-Driven Exploration: Challenges

- It is difficult for domain experts to explore data
  - Dependency on data scientists distances domain experts from the data
  - Analyses are mostly confirmatory (Tukey, 1977) – batch-oriented analysis hampers exploration
- Exploratory tasks are inherently iterative as one tests and formulates hypotheses



[Modified from Van Wijk, Vis 2005]

NYU TANDON SCHOOL OF ENGINEERING

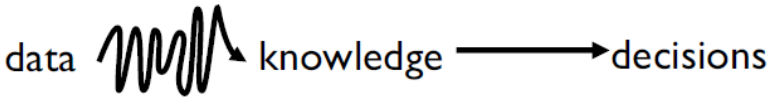
VIDA VISUALIZATION IMAGING AND DATA ANALYSIS CENTER

J. Freire, 2021

Data-Driven Exploration: Challenges

- After many steps…
  - It is easy to get lost and not remember how a result was derived
  - Processes can break or misbehave in unforeseen ways
  - Results can be hard to understand, interpret and trust

"An analysis has 30 different steps. It is tempting to just do this then that and then this. You have no idea in which ways you are wrong and what data is wrong" [Kandel et al., VAST 2012]



Incorrect conclusions can have serious consequences

NYU TANDON SCHOOL OF ENGINEERING

VIDA VISUALIZATION IMAGING AND DATA ANALYSIS CENTER

Processes, logging, reproducibility

- Processes are everywhere
  - Consisting of multiple steps
  - Reading and creating data
- Process engines are good at running processes quickly
  - SQL engines, workflow engines, BPM engines, ETL engines, …
- Logging usually only implemented for debugging
  - Technical focus
- Such log files cannot easily explain why something was computed (or not)
- Provenance and explainability
- Provenance and reproducibility

Definitions

- Oxford English Dictionary:
  - the fact of coming from some particular source or quarter; origin, derivation
  - the history or pedigree of a work of art, manuscript, rare book, etc.;
  - concretely, a record of the passage of an item through its various owners.
- The provenance of a piece of data is the process that led to that piece of data

Provenance in a Hospital

- Imagine we would record everything that is done with a patient when in a clinic …

- Do we treat all patients with same diagnosis the same?
- Do we treat patients according to clinical guidelines?
- When we treat patients differently – are there groups?
- What is different between such groups?
- Do patients we treat differently have different outcome?
- Does the same treatment take the same time?
- …

Example

Provenienz

- "Provenienz (von lateinisch *provenire* „herkommen") bezeichnet allgemein die Herkunft einer Person oder Sache.[1] Besondere Bedeutung hat der Begriff als Bezeichnung der Herkunft von Kunstwerken und Kulturgütern, ihrer Erforschung widmet sich die Provenienzforschung. Der Begriff ist auch als Herkunftsangabe von Waren geläufig, meist im Sinne einer Qualitätsangabe. „
  - Wikipedia, 10/21

Management and Analysis of Provence Data

- Provenance itself is data
  - Logfiles, lineage, monitoring files, audit trails, …
- These must be managed
  - Stored: Efficiently, compressed, quickly, …
  - Searched: Keywords, structure, queries, …
  - Analyzed: Aggregated, error detection, pattern matching, …

Who should be here

- Bachelor Informatik / IMP / Kombi / Infomit
- Ability to read English papers
- Knowledge in
  - Data management (e.g. databases, data models)
  - Systems Engineering (components, monitoring, …)
- Willingness to independent work
  - Search suitable papers covering a topic, prepare presentations, write seminar thesis

How it will work

- Today: Presentation and choice of topics
  - If desired, we will group teams of 2 students
- 21.11.2021: Send an outline of your topic (next slide)
- ~17.12.21: Present your topic in 5min flash-presentation
- ~15.1.22: Meet your advisor to discuss slides
- ~10.2.22: Present your topic (30min) in a Blockseminar
- 30.3.2022: Write seminar thesis (10-15 pages)

The outline

- Topics are rather abstract
- Find a set of suitable papers covering the topic
  - Screen 20-30, read 5-8, focus on 2-3
  - Focus is allowed and welcome
- Extract the most important information
- Structure into an outline of your seminar paper
  - Chapters, sections, 1-2 sentences per section to describe the content, 2-4 figures
- Write an abstract
  - Roughly 30 lines – what is the topic, what will the thesis describe?
- Send me abstract + outline + paper references
  - Mark your top-3 references – those that most likely will form the basis of your work

The 5-min flash talk

- Focus on marketing – attract students to your topic
  - What is the topic?
  - Why is it challenging?
  - Why is it cool?
  - What are important applications?
  - What will your talk be about?
- At most 5 slides
- Focus on figures & examples
  - Omit details, formulas or algorithms

Presentation

- 30min presentation
- German or English
- Explain topic, methods, experimental results
- Compare different approaches
- Aim: Your audience should understand what you say
- No need to cover the literature entirely – select most important contríbutions

Teams

- If a topic is addressed by a team of two students, I expect
  - Read more papers
  - Have more content in your outline
  - Be more exhaustive and detailed in your thesis
  - Presentations times remain the same – choose wisely

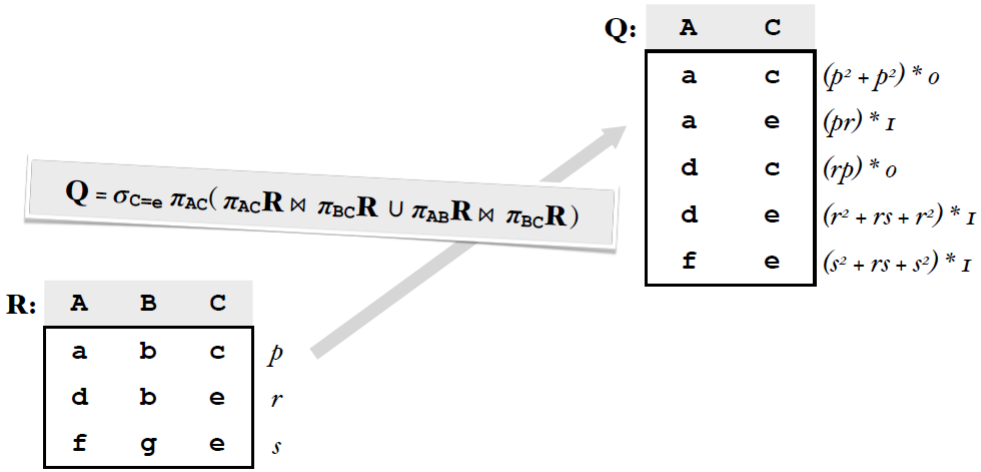
ToC

- Introduction
- Topics
- Assignment
- Hints on presenting your topic and writing your thesis

Topic	Assigned to
Provenance Semi-Rings	
Why and Why not provenance	
Tools zum Logfile Management	
Anfragesprachen für Provenncedaten	
Provenance-Modelle und Standards	
Effizientes Provenance-Management / Storage	
Provenance in Scientific Workflow Systems	
Process Mining - conformance checking	
Process Mining - process recovery	
Visualisierung von Provenance-Daten	
Provenance für Reproducibility	
YOUR OWN TOPIC HERE	

Provenance Semirings

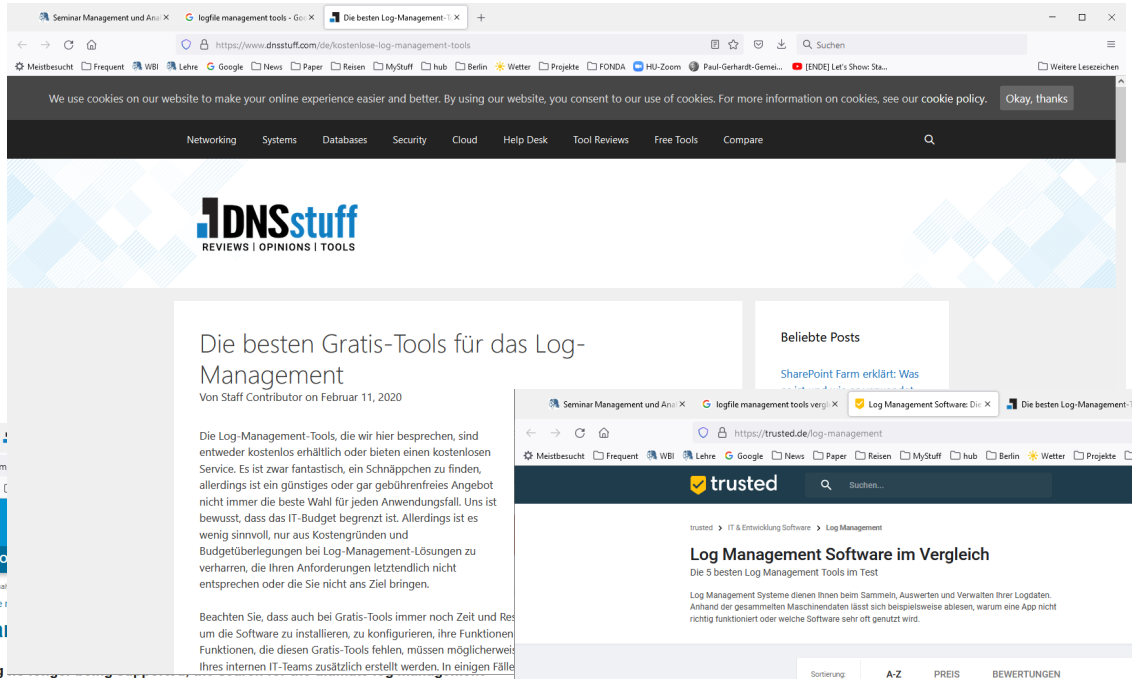
- Semirings: A particular algebraic structure
- Represent query processing as polynomials over semirings
- Allows fast computation of "why provenance" – why is a tuple in the result of a relational query?

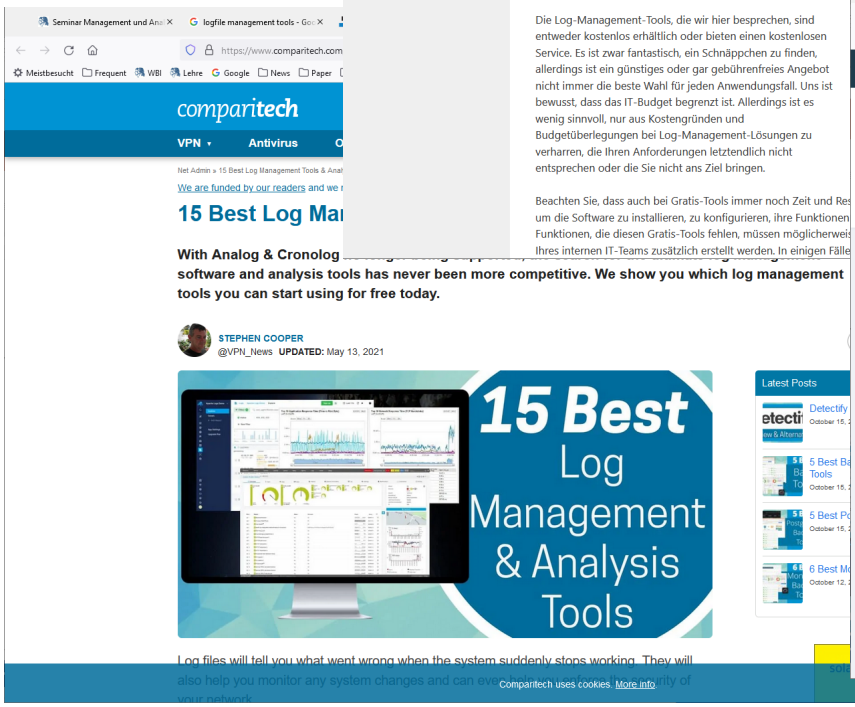


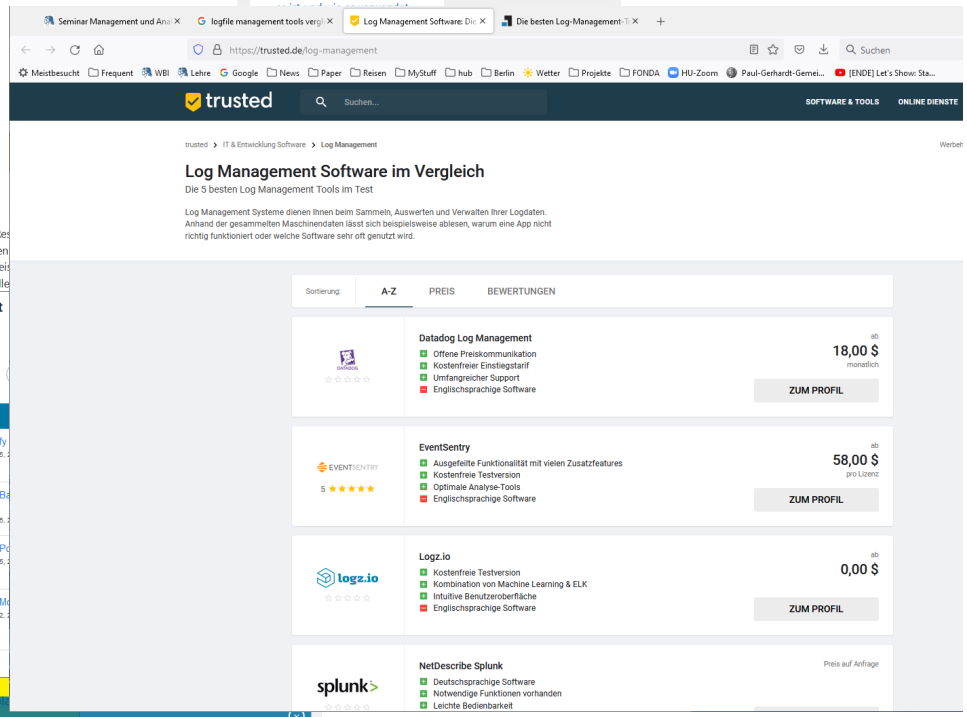
Why and Why not provenance

- Debugging a query often also involves missing tuples
  - Should be there
  - But isn't
  - Why not?
- "Why not" provenance – why a given tuple is not in the result of a query

Tools zum Logfile Management







Anfragesprachen für Provenancedaten

- Give me all log entries of the form …
- Give me all entries following an entry like …
- Count the number of entries like …
- Group all entries by this pattern and count occurrences
- Return the part of the log file between ? And ?
- …
- Queries over (structured, time stamped) log files
- Queries must adhere to a query language

Provenance-Modelle und Standards

- Tools, query languages, indexing schemes …
- … would benefit when all log entries would look the same
  - Same attributes, same things named the same, …
- But provenance models differ
  - Subentries? Query or process engine? …
- What are standards for modeling and managing provenance data?

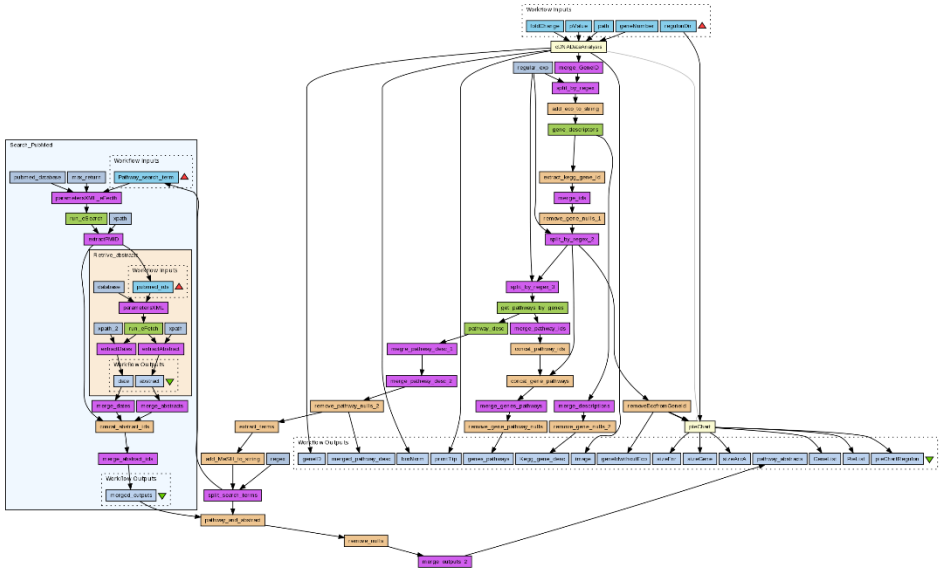
Effizientes Provenance-Management / Storage

- Provenance data can become pretty huge
- While a process is running, it flows in rapidly
- Sometimes, we want an analysis of provenance data while the process is still running – streaming
- Provenance in distributed systems must be efficiently collected and made available at a single location

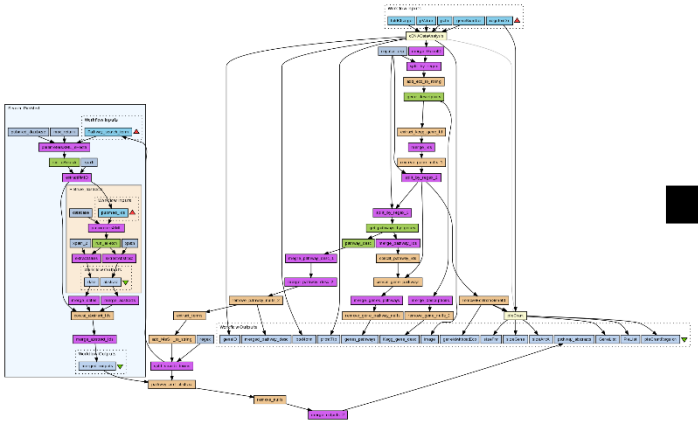
- How can we manage provenance data efficiently?

Provenance in Scientific Workflow Systems

- Scientific workflows – programming in the large for scientific data analysis
- Provenance for
  - Debugging – what went wrong?
  - Certification – what did I do?
  - Reproducibility – look what I have done
  - …
- Applications, models, storage, queries, visualization, …



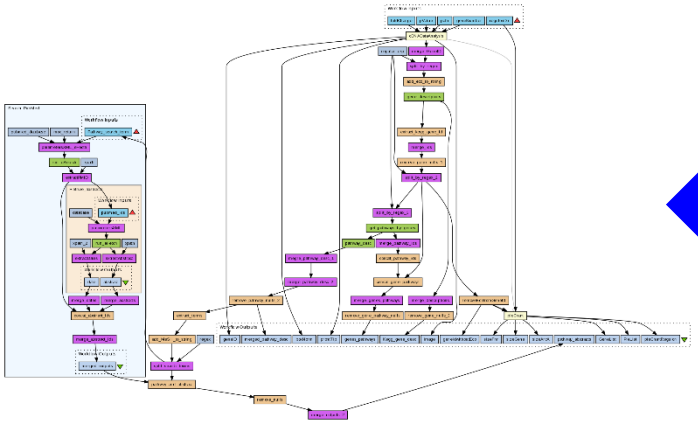
Process Mining - Conformance checking



#	IP	Id	Acces	Time	Method/URL/Protocol	Status	Bytes	Referer	Agent
1	165.182.168.101	-	-	16/06/2002:16:24:06	GET p1.htm HTTP/1.1	200	3821	out.htm	Mozilla/4.0 (MSIE 5.5; WinNT 5.1)
2	165.182.168.101	-	-	16/06/2002:16:24:10	GET A.gif HTTP/1.1	200	3766	p1.htm	Mozilla/4.0 (MSIE 5.5; WinNT 5.1)
3	165.182.168.101	-	-	16/06/2002:16:24:57	GET B.gif HTTP/1.1	200	2878	p1.htm	Mozilla/4.0 (MSIE 5.5; WinNT 5.1)
4	204.231.180.195	-	-	16/06/2002:16:32:06	GET p3.htm HTTP/1.1	304	0	-	Mozilla/4.0 (MSIE 6.0; Win98)
5	204.231.180.195	-	-	16/06/2002:16:32:20	GET C.gif HTTP/1.1	304	0	-	Mozilla/4.0 (MSIE 6.0; Win98)
6	204.231.180.195	-	-	16/06/2002:16:34:10	GET p1.htm HTTP/1.1	200	3821	p3.htm	Mozilla/4.0 (MSIE 6.0; Win98)
7	204.231.180.195	-	-	16/06/2002:16:34:31	GET A.gif HTTP/1.1	200	3766	p1.htm	Mozilla/4.0 (MSIE 6.0; Win98)
8	204.231.180.195	-	-	16/06/2002:16:34:53	GET B.gif HTTP/1.1	200	2878	p1.htm	Mozilla/4.0 (MSIE 6.0; Win98)
9	204.231.180.195	-	-	16/06/2002:16:38:40	GET p2.htm HTTP/1.1	200	2960	p1.htm	Mozilla/4.0 (MSIE 6.0; Win98)
10	165.182.168.101	-	-	16/06/2002:16:39:02	GET p1.htm HTTP/1.1	200	3821	out.htm	Mozilla/4.0 (MSIE 5.01; WinNT 5.1)
11	165.182.168.101	-	-	16/06/2002:16:39:15	GET A.gif HTTP/1.1	200	3766	p1.htm	Mozilla/4.0 (MSIE 5.01; WinNT 5.1)
12	165.182.168.101	-	-	16/06/2002:16:39:45	GET B.gif HTTP/1.1	200	2878	p1.htm	Mozilla/4.0 (MSIE 5.01; WinNT 5.1)
13	165.182.168.101	-	-	16/06/2002:16:39:58	GET p2.htm HTTP/1.1	200	2960	p1.htm	Mozilla/4.0 (MSIE 5.01; WinNT 5.1)
14	165.182.168.101	-	-	16/06/2002:16:42:03	GET p3.htm HTTP/1.1	200	4036	p2.htm	Mozilla/4.0 (MSIE 5.01; WinNT 5.1)
15	165.182.168.101	-	-	16/06/2002:16:42:07	GET p2.htm HTTP/1.1	200	2960	p1.htm	Mozilla/4.0 (MSIE 5.5; WinNT 5.1)
16	165.182.168.101	-	-	16/06/2002:16:42:08	GET C.gif HTTP/1.1	200	3423	p2.htm	Mozilla/4.0 (MSIE 5.01; WinNT 5.1)
17	204.231.180.195	-	-	16/06/2002:17:34:20	GET p3.htm HTTP/1.1	200	2342	out.htm	Mozilla/4.0 (MSIE 6.0; Win98)
18	204.231.180.195	-	-	16/06/2002:17:34:48	GET C.gif HTTP/1.1	200	3423	p2.htm	Mozilla/4.0 (MSIE 6.0; Win98)
19	204.231.180.195	-	-	16/06/2002:17:35:45	GET p4.htm HTTP/1.1	200	3523	p3.htm	Mozilla/4.0 (MSIE 6.0; Win98)
20	204.231.180.195	-	-	16/06/2002:17:35:56	GET D.gif HTTP/1.1	200	3231	p4.htm	Mozilla/4.0 (MSIE 6.0; Win98)
21	204.231.180.195	-	-	16/06/2002:17:36:06	GET E.gif HTTP/1.1	404	0	p4.htm	Mozilla/4.0 (MSIE 6.0; Win98)

- If the process specification is a law (business rule) –
- … then provenance data is a proof of compliance
- Conformance checking: Did a process, as represented in a log, conform to the specification?
  - Big issues when "humans are in the loop"
- Of course: Complexity depends on complexity of specification language

Process Mining - process recovery



#	IP	Id	Acces	Time	Method/URL/Protocol	Status	Bytes	Referer	Agent
1	165.182.168.101	-	-	16/06/2002:16:24:06	GET p1.htm HTTP/1.1	200	3821	out.htm	Mozilla/4.0 (MSIE 5.5; WinNT 5.1)
2	165.182.168.101	-	-	16/06/2002:16:24:10	GET A.gif HTTP/1.1	200	3766	p1.htm	Mozilla/4.0 (MSIE 5.5; WinNT 5.1)
3	165.182.168.101	-	-	16/06/2002:16:24:57	GET B.gif HTTP/1.1	200	2878	p1.htm	Mozilla/4.0 (MSIE 5.5; WinNT 5.1)
4	204.231.180.195	-	-	16/06/2002:16:32:06	GET p3.htm HTTP/1.1	304	0	-	Mozilla/4.0 (MSIE 6.0; Win98)
5	204.231.180.195	-	-	16/06/2002:16:32:20	GET C.gif HTTP/1.1	304	0	-	Mozilla/4.0 (MSIE 6.0; Win98)
6	204.231.180.195	-	-	16/06/2002:16:34:10	GET p1.htm HTTP/1.1	200	3821	p3.htm	Mozilla/4.0 (MSIE 6.0; Win98)
7	204.231.180.195	-	-	16/06/2002:16:34:31	GET A.gif HTTP/1.1	200	3766	p1.htm	Mozilla/4.0 (MSIE 6.0; Win98)
8	204.231.180.195	-	-	16/06/2002:16:34:53	GET B.gif HTTP/1.1	200	2878	p1.htm	Mozilla/4.0 (MSIE 6.0; Win98)
9	204.231.180.195	-	-	16/06/2002:16:38:40	GET p2.htm HTTP/1.1	200	2960	p1.htm	Mozilla/4.0 (MSIE 6.0; Win98)
10	165.182.168.101	-	-	16/06/2002:16:39:02	GET p1.htm HTTP/1.1	200	3821	out.htm	Mozilla/4.0 (MSIE 5.01; WinNT 5.1)
11	165.182.168.101	-	-	16/06/2002:16:39:15	GET A.gif HTTP/1.1	200	3766	p1.htm	Mozilla/4.0 (MSIE 5.01; WinNT 5.1)
12	165.182.168.101	-	-	16/06/2002:16:39:45	GET B.gif HTTP/1.1	200	2878	p1.htm	Mozilla/4.0 (MSIE 5.01; WinNT 5.1)
13	165.182.168.101	-	-	16/06/2002:16:39:58	GET p2.htm HTTP/1.1	200	2960	p1.htm	Mozilla/4.0 (MSIE 5.01; WinNT 5.1)
14	165.182.168.101	-	-	16/06/2002:16:42:03	GET p3.htm HTTP/1.1	200	4036	p2.htm	Mozilla/4.0 (MSIE 5.01; WinNT 5.1)
15	165.182.168.101	-	-	16/06/2002:16:42:07	GET p2.htm HTTP/1.1	200	2960	p1.htm	Mozilla/4.0 (MSIE 5.5; WinNT 5.1)
16	165.182.168.101	-	-	16/06/2002:16:42:08	GET C.gif HTTP/1.1	200	3423	p2.htm	Mozilla/4.0 (MSIE 5.01; WinNT 5.1)
17	204.231.180.195	-	-	16/06/2002:17:34:20	GET p3.htm HTTP/1.1	200	2342	out.htm	Mozilla/4.0 (MSIE 6.0; Win98)
18	204.231.180.195	-	-	16/06/2002:17:34:48	GET C.gif HTTP/1.1	200	3423	p2.htm	Mozilla/4.0 (MSIE 6.0; Win98)
19	204.231.180.195	-	-	16/06/2002:17:35:45	GET p4.htm HTTP/1.1	200	3523	p3.htm	Mozilla/4.0 (MSIE 6.0; Win98)
20	204.231.180.195	-	-	16/06/2002:17:35:56	GET D.gif HTTP/1.1	200	3231	p4.htm	Mozilla/4.0 (MSIE 6.0; Win98)
21	204.231.180.195	-	-	16/06/2002:17:36:06	GET E.gif HTTP/1.1	404	0	p4.htm	Mozilla/4.0 (MSIE 6.0; Win98)

- Sometimes we have no specification but only a log
- Compute "the" process specification that could have produced this log
  - The smallest? The most likely? With/-out outliers (e.g. errors)?

Visualisierung von Provenance-Daten

- Provenance data can be extremely large
- Visualization: Allow intuitive exploration by humans
  - Proper aggregation of entries
  - Layout and clustering of entries
  - Visualization including the specification
  - …

Provenance für Reproducibility

- Reproducibility crisis – many published empirical scientific results cannot be reproduced
  - Results deduced from such experiments might be plain wrong
  - Across all subjects, including computer science
- Provenance is an aid for computational reproducibility
  - Can I proof what I computed?
  - Can I repeat the computation?
  - Can I repeat the computation on a different machine?
  - …
- Reusability, reproducibility, replicability, …

ToC

- Introduction
- Topics
- Assignment
- Hints on presenting your topic and writing your thesis

Allgemeine Hinweise

- Dozenten sind ansprechbar!
  - Vorbesprechung des Themas
  - Folien durchgehen
  - Abgrenzung der Ausarbeitung
- Diskussion erwünscht
  - Keine Angst vor Fragen: Fragen sind keine Kritik
  - Eine Frage nicht beantworten können ist in Ordnung
- Tiefe, nicht Breite
  - Lieber das Thema einengen und dafür Details erklären
- Bezug nehmen
  - Vergleich zu anderen Arbeiten (im Seminar)

Allgemeine Hinweise

- Werten und bewerten
  - Keine Angst vor nicht ganz zutreffenden Aussagen – solange gute Gründe vorhanden sind
  - Begründen und argumentieren
  - Kritikloses Abschreiben ist fehl am Platz
- Literaturrecherche ist notwendig
  - Die ausgegebenen Arbeiten sind Anker
  - Weiterführende Arbeiten müssen herangezogen werden
  - Auch Grundlagen nachlesen

- Wir schicken eine Liste zum Abhaken rum

Wie halte ich einen Seminarvortrag

1. Wenn man nun so einen Seminarvortrag halten muss, dann empfiehlt es sich, möglichst lange Sätze auf die Folien zu schreiben, damit die Zuhörer nach dem Vortrag aus den Folienkopien noch wissen, was man eigentlich gesagt hat.

2. Während so einem Vortrag schaut sowieso jeder zum Projektor, also kann man das selbst ruhig auch tun - damit kontrolliert man gleichzeitig auch, ob der Beamer wirklich alles projeziert, was auf dem Laptop zu sehen ist. Ausserdem kann man so den Strom für das Laptop-Display sparen.

3. Übersichtsfolien am Anfang sind langweilig, enthalten keinen Inhalt und nehmen den Zuhörern die ganze Spannung. Schliesslich gibt‘s im Kino am Anfang auch keine Inhaltsangabe.

4. Powerpoint kann viele lustige Effekte, hat tolle Designs und Animationen. Die sollte man zur Auflockerung des Vortrags unbedingt alle benutzen, um zu zeigen, wie gut man das Tool im Griff hat.

5. Nicht zu wenig auf die Folien schreiben. Man weiß ja nie, ob man sie nicht doch ausdrucken muss, und man kann so wertvolle Zeit sparen, wenn man nicht weiterschalten muss.

6. Man sollte versuchen, möglichst lange zu reden. Die Zeitvorgaben sind nur für die Leute, die nicht genug wissen - eigentlich will der Prüfer sehen, dass man sich auch darüber hinaus mit dem Thema beschäftigt hat.

Bloß keine Hervorhebungen im Text – sonst müssen die Zuhörer ja gar nicht mehr aufpassen!

Hinweise zum Vortrag

- ~30 Minuten inkl Diskussion
- Klare Gliederung
- Ab und an Hinweise geben, wo man sich befindet
- Bilder und Grafiken; Beispiele
- Font: mind. 16pt
- Eher Stichwörter als lange Sätze
- Vorträge können auch unterhaltend sein
  - Gimmicks, Rhythmuswechsel, Einbeziehen der Zuhörer, etc.
- Adressat sind alle Teilnehmer, nicht nur die Betreuer
- Technik: Laptop? Powerpoint?

Hinweise zur Ausarbeitung

- Eine gedruckte Version abgeben
  - Selbstständigkeitserklärung unterschreiben
- Eine elektronische Version schicken
- Referenzen: Alle verwendeten und nur die
  - Im Text referenzieren, Liste am Schluss
- Korrekt zitieren
  - Vorsicht vor Übernahme von kompletten Textpassagen; wenn, dann deutlich kennzeichnen
  - Aussagen mit Evidenz oder Verweis auf Literatur versehen
- Verwendung von gefundenen Arbeiten im Web
  - Möglich, aber VORSICHT
  - Eventuell Themenschwerpunkt verschieben – Betreuer fragen

Format

- Benutzung unserer Latex-Vorlage
- Nur eine Schriftart, wenig und konsistente Wechsel in Schriftgröße und –stärke
- Inhaltsverzeichnis
- Bilder: Nummerieren und darauf verweisen
- Referenzen:
  - [1] Yan, X., Yu, P. S. and Han, J. (2004). "Graph Indexing: A Frequent Structure-Based Approach". SIGMOD, Paris, France.
  - [YYH04] Yan, X., Yu, P. S. and Han, J. (2004). "Graph Indexing: A Frequent Structure-Based Approach". SIGMOD, Paris, France.
- Darf man Wikipedia zitieren?
  - Ja, aber nicht dauernd

Hinweise zur Ausarbeitung –2-

- Gezielt und sachlich schreiben
  - Ausführungen zur „Philosophische Überlegungen zu Vorzügen probabilistischer Verfahren im Vergleich zu Dempster‘s Theory of Evidence“ oder zur „Anmerkungen zur Trivialisierung des politischen Diskurs für soziale Netzwerke unter besonderer Berücksichtigung von Twitter“ möglichst kurz halten
  - Füllwörter vermeiden (dabei, hierbei, dann, …)
  - Knappe Darlegung, präzise Sprache
- Eine gute Gliederung ist die halbe Miete
- Kommen Sie zu Aussagen
  - Vorteile, Nachteile, verwandte Arbeiten, mögliche Erweiterungen, Anwendbarkeit, eigene Erfahrungen, …