12 July 2019

EUROPEAN
COURT
OF AUDITORS

Process Mining in the Wild

Zsolt VARGA, ECALab

zsolt.varga@eca.europa.eu

Process Mining Safari

- Case study 1: process mining as a new form of data representation
    - EU public consultations, a relatively simple case

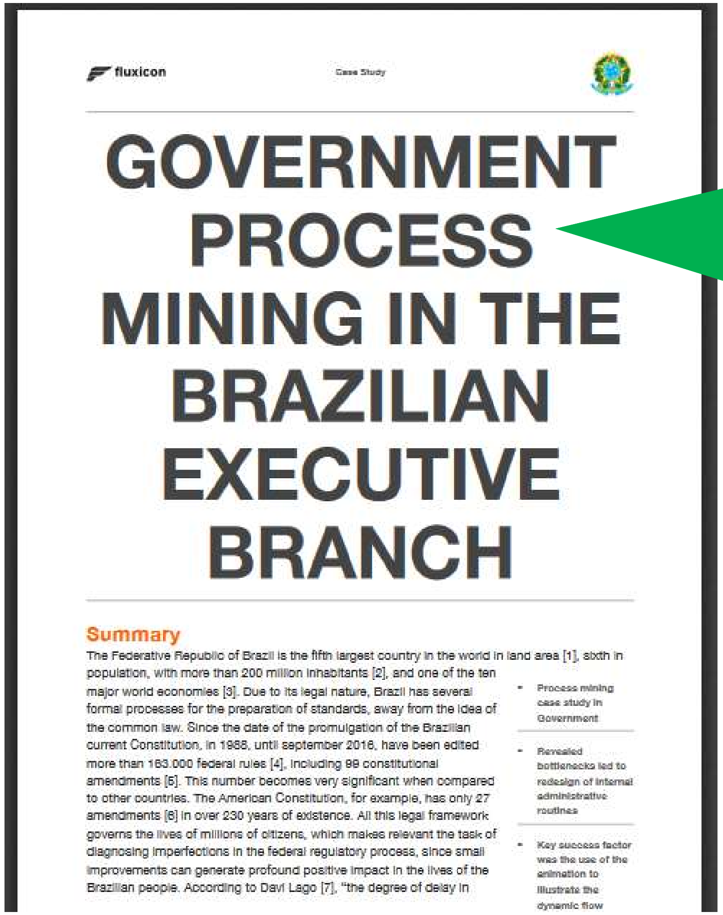
- Case study 2: process mining for conformance checking
    - Audit of EU agencies, a bit more complicated
        - Getting the data, extracting from pdf files, data protection
        - Adding calculated fields in the event log
        - Fluxicon disco vs. Prom6

- Case study 3: exploratory process mining in „big data”
    - CA Service Desk manager database tables
        - Text mining methods for making sense of process data

Process mining in the government sector

https://fluxicon.com/blog/2018/12/case-study-government-process-mining-in-the-brazilian-executive-branch/

http://romualdoalves.com/decretos-2015-a-2018-poc-text-mining/



fluxicon   Case Study

GOVERNMENT PROCESS MINING IN THE BRAZILIAN EXECUTIVE BRANCH

Summary

The Federative Republic of Brazil is the fifth largest country in the world in land area [1], sixth in population, with more than 200 million inhabitants [2], and one of the ten major world economies [3]. Due to its legal nature, Brazil has several formal processes for the preparation of standards, away from the idea of the common law. Since the date of the promulgation of the Brazilian current Constitution, in 1988, until september 2016, have been edited more than 163.000 federal rules [4], including 99 constitutional amendments [5]. This number becomes very significant when compared to other countries. The American Constitution, for example, has only 27 amendments [6] in over 230 years of existence. All this legal framework governs the lives of millions of citizens, which makes relevant the task of diagnosing imperfections in the federal regulatory process, since small improvements can generate profound positive impact in the lives of the Brazilian people. According to Davi Lago [7], "the degree of delay in

- Process mining case study in Government
- Revealed bottlenecks led to redesign of internal administrative routines
- Key success factor was the use of the animation to illustrate the dynamic flow

Inspiration for using process mining in performance audit. It made me try PM on the audit on public consultations.

Process data is everywhere!

Process Mining

for performance audit

Case study 1:

European Public Consultations

Process Mining Safari Part 1: playing it safe

Audit case

- Audit question: How are public consultations implemented by the Commission?
- Dataset: Excel file about the various steps of public consultations and their start dates
- Business need: visualise the relative durations of PCs compared to each other and show deviations from the „ideal” process path

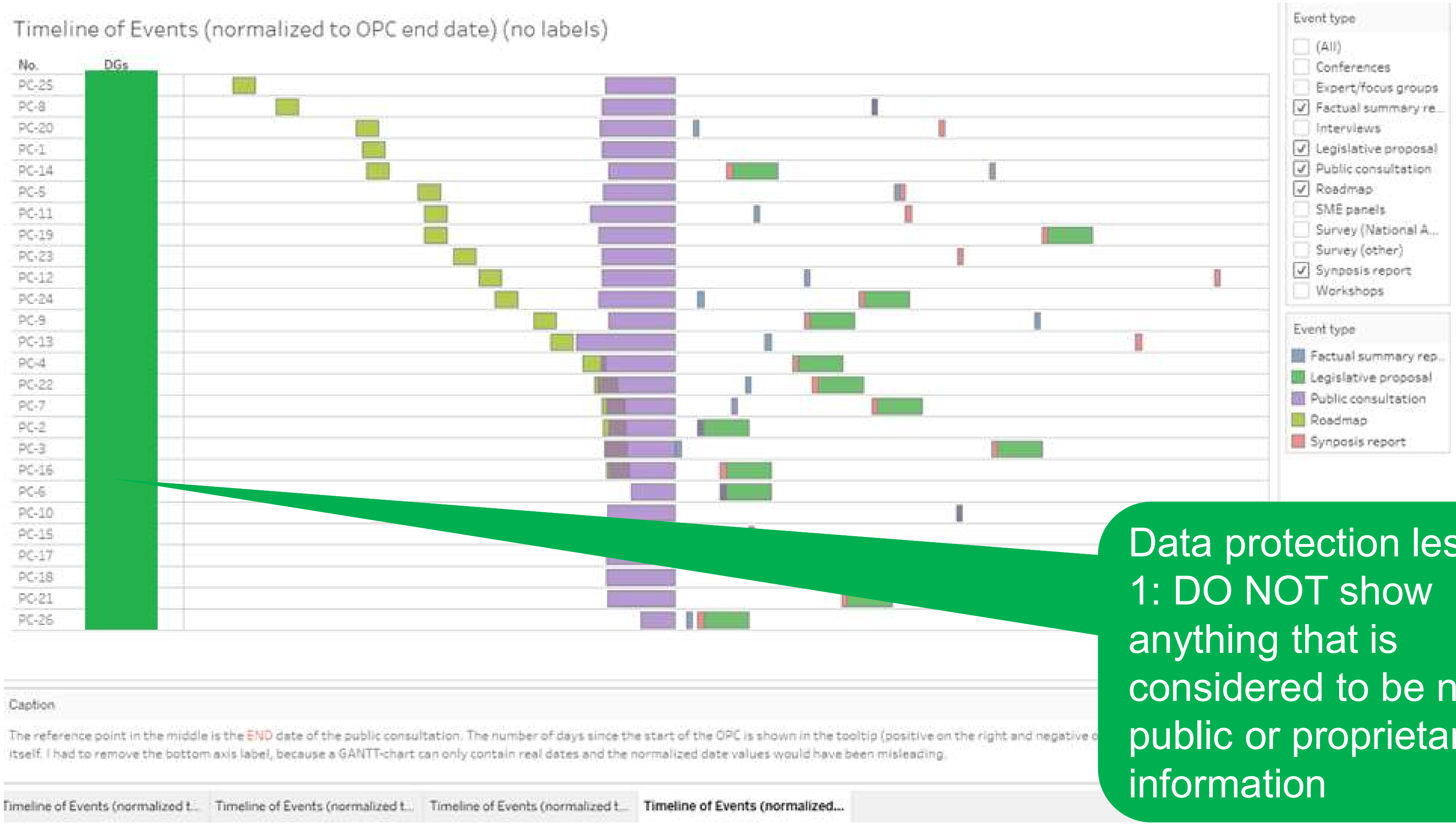
Input data for process mining

Cons_ID	Event	Start	End_theoretical
PC-10	Survey (National Authorities)	2014-09-01	2015-02-01
PC-10	Conferences	2014-11-01	2016-03-01
PC-10	Expert/focus groups	2015-03-01	2016-02-01
PC-1	Roadmap	2015-05-07	2015-06-04
PC-18	Survey (other)	2015-06-26	2015-09-01
PC-11	Roadmap	2015-07-23	2015-08-20
PC-20	Roadmap	2015-09-28	2015-10-26
PC-11	Expert/focus groups	2015-10-01	2016-05-01
PC-10	Workshops	2015-10-01	2015-10-31
PC-5	Roadmap	2015-11-30	2015-12-28
PC-12	Roadmap	2015-12-08	2016-01-05
PC-19	Roadmap	2015-12-08	2016-01-05
PC-10	OPC	2016-01-12	2016-04-0
PC-8	Roadmap	2016-01-18	20
PC-9	Conferences	2016-02-01	[illegible]
PC-11	OPC	2016-02-1	-31
PC-11	Survey (other)	2	2016-05-31
PC-1	OPC	2016-03-01	2016-06-01
PC-20	Interviews	2016-03-15	2016-08-24
PC-14	Expert/focus groups	2016-04-01	2017-04-01

Open public consultations: events and start/completion dates for each event within an OPC.

Process data is everywhere!

Tableau viz, pretty close, but static, concurrencies shown as overlaps


Timeline of Events (normalized to OPC end date) (no labels)
No.
DGs
PC-25
PC-8
PC-20
PC-1
PC-14
PC-5
PC-11
PC-19
PC-23
PC-12
PC-24
PC-9
PC-13
PC-4
PC-22
PC-7
PC-2
PC-3
PC-16
PC-6
PC-10
PC-15
PC-17
PC-18
PC-21
PC-26
Timeline of Events (normalized t...
Timeline of Events (normalized t...
Timeline of Events (normalized t...
Timeline of Events (normalized...

Event type

☐ (All)
☐ Conferences
☐ Expert/focus groups
☑ Factual summary re...
☐ Interviews
☑ Legislative proposal
☑ Public consultation
☑ Roadmap
☐ SME panels
☐ Survey (National A...
☐ Survey (other)
☑ Synposis report
☐ Workshops

Event type

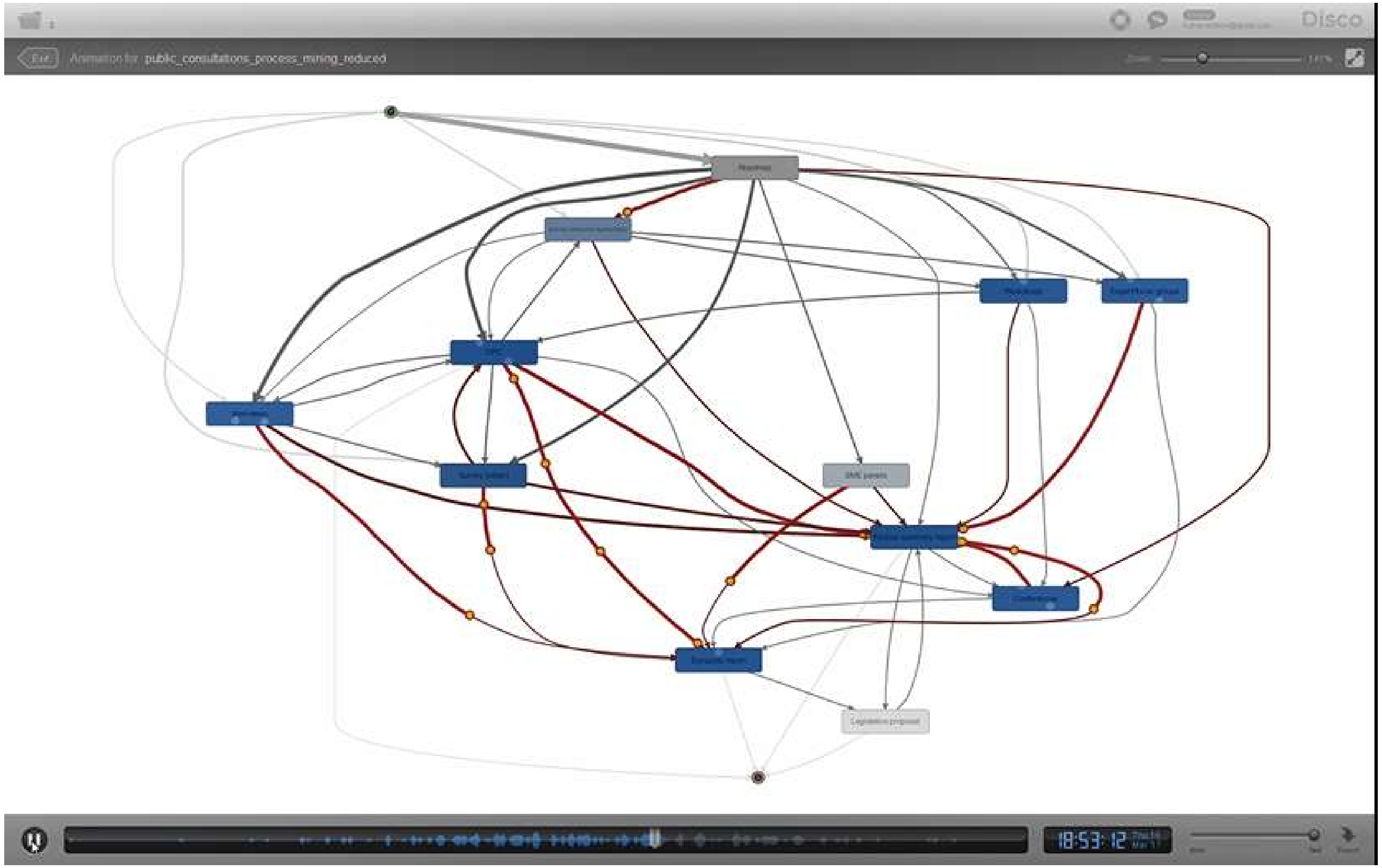
- 🟦 Factual summary rep...
- 🟩 Legislative proposal
- 🟪 Public consultation
- 🟩 Roadmap
- 🟥 Synposis report

Caption

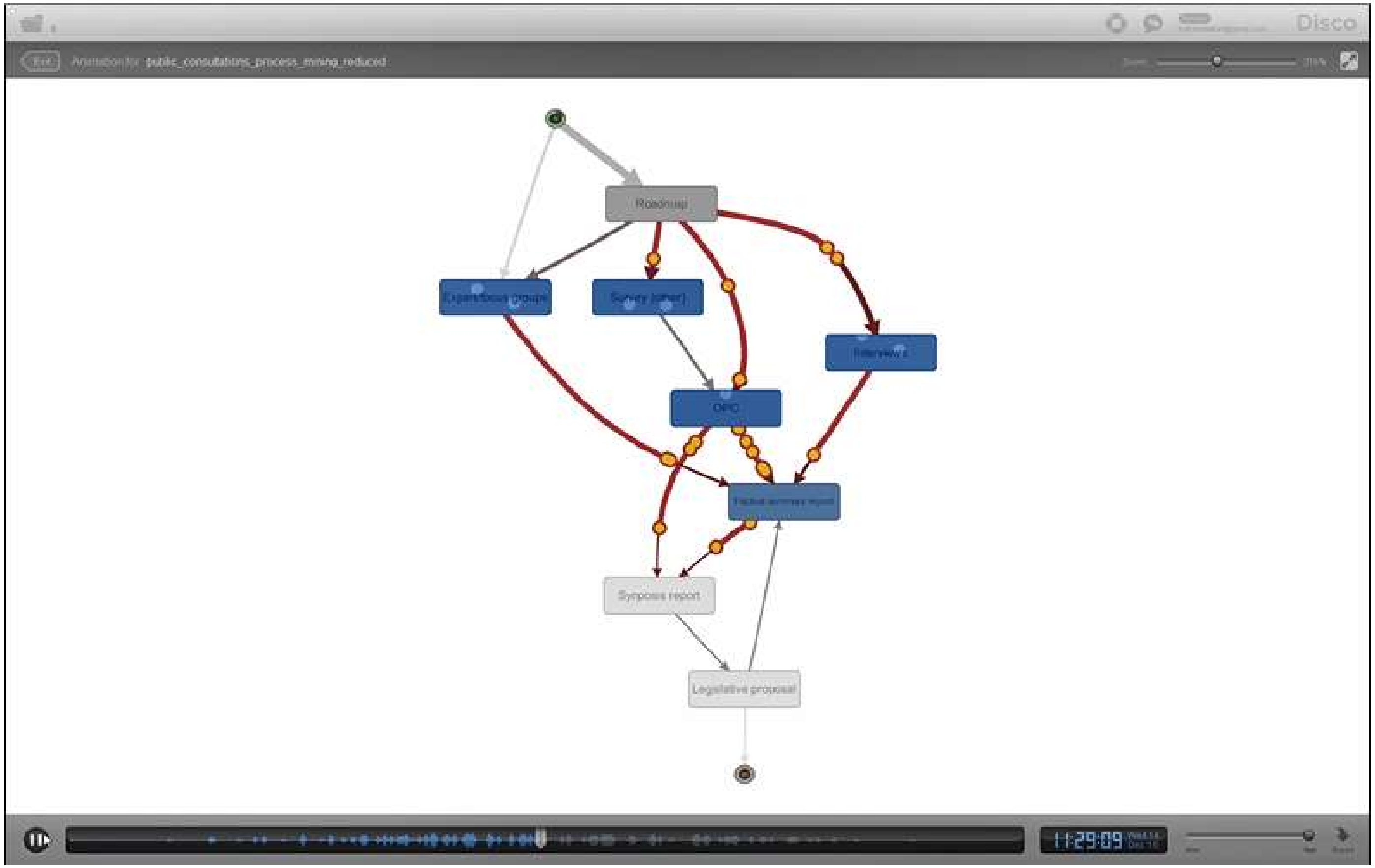
The reference point in the middle is the END date of the public consultation. The number of days since the start of the OPC is shown in the tooltip (positive on the right and negative o
itself. I had to remove the bottom axis label, because a GANTT-chart can only contain real dates and the normalized date values would have been misleading.

Data protection lesson 1: DO NOT show anything that is considered to be non-public or proprietary information

Full OPC process (spaghetti)

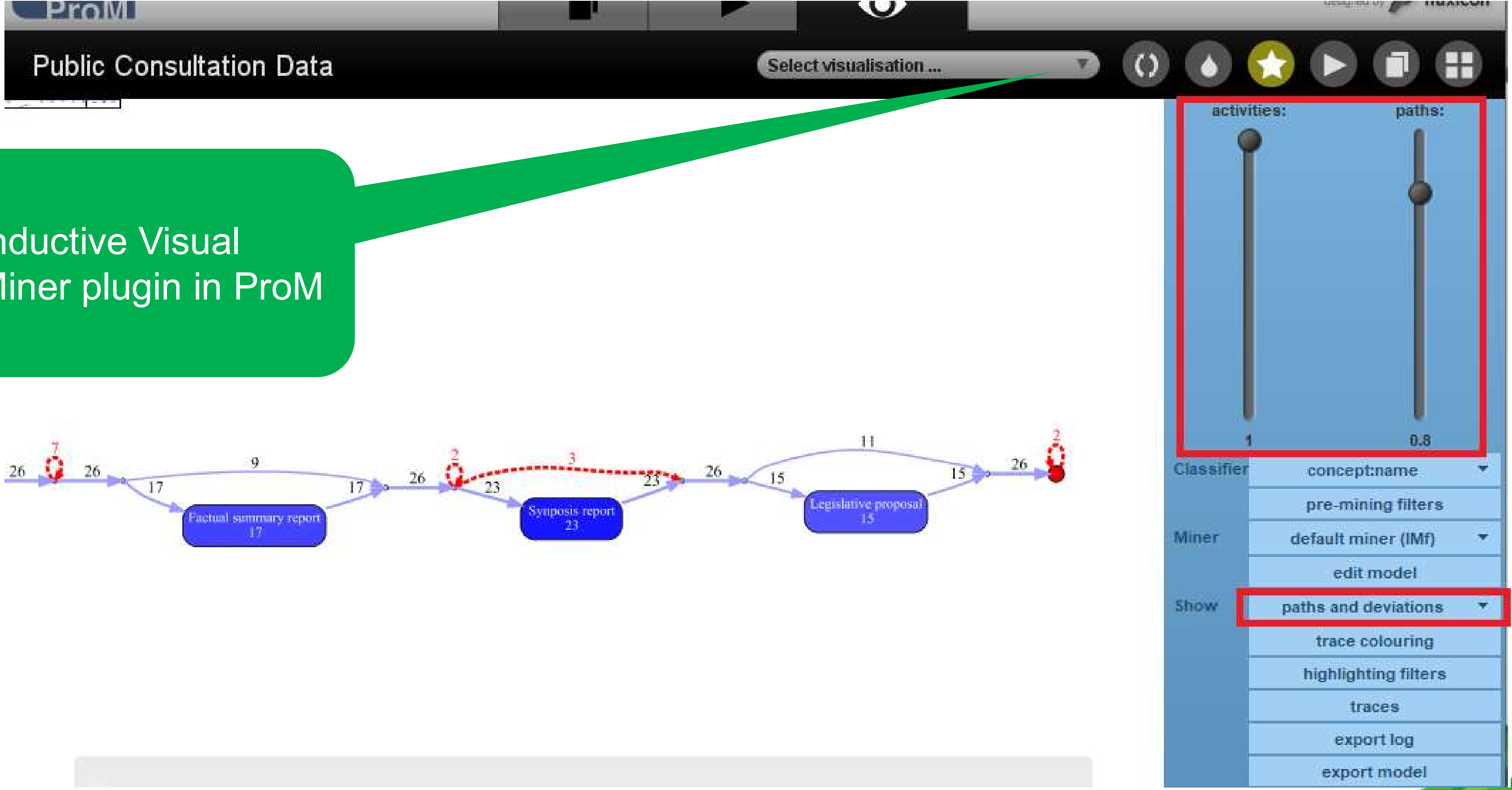


Simplified OPC process (layered, „lasagne” process)



Process conformance checking on automatically inferred model

Inductive Visual Miner plugin in ProM



Public Consultation Data

activities: 1

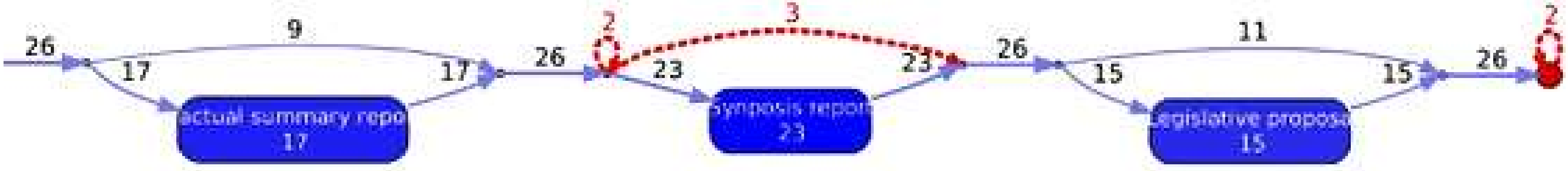
paths: 0.8

Classifier: concept:name

Miner: default miner (IMf)

Show: paths and deviations

Process conformance checking on automatically inferred model



3 OPCs didn’t have a synopsis report

2 OPCs had additional steps after the legislative proposal

Lessons learned

- There’s process data everywhere
- ProM/Disco is suitable for visualising performance audit results
- Replaying the log on the process model adds the time dimension by animating the path of the traces in the model
- A great way to communicate complex ideas quickly. When I demonstrated it to a visiting delegation, they immediately spotted the fact that two consultations didn’t finish at the legal proposal (no need for prior briefing).

Audit of EU agencies

Visualising and conformance checking of agency payments

Process Mining Safari Part 2: a closer look at the beast

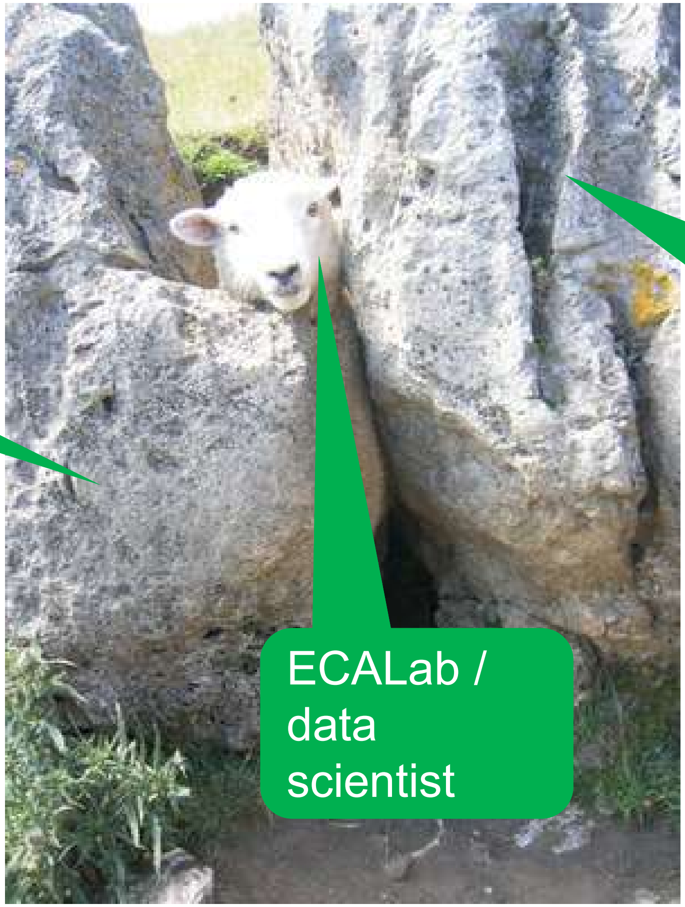
Audit case

- Audit question: How can we automate the audit of EU agencies' payment processes?

- Dataset: the EC's accounting system has all the required data for certain types of agencies, but there are obstacles to directly getting the data from the database tables

- Business need: visual representation of the process workflow highlighting process deviation and non-compliance with duty segregation and deadlines

Small detour (between rock and a hard place)

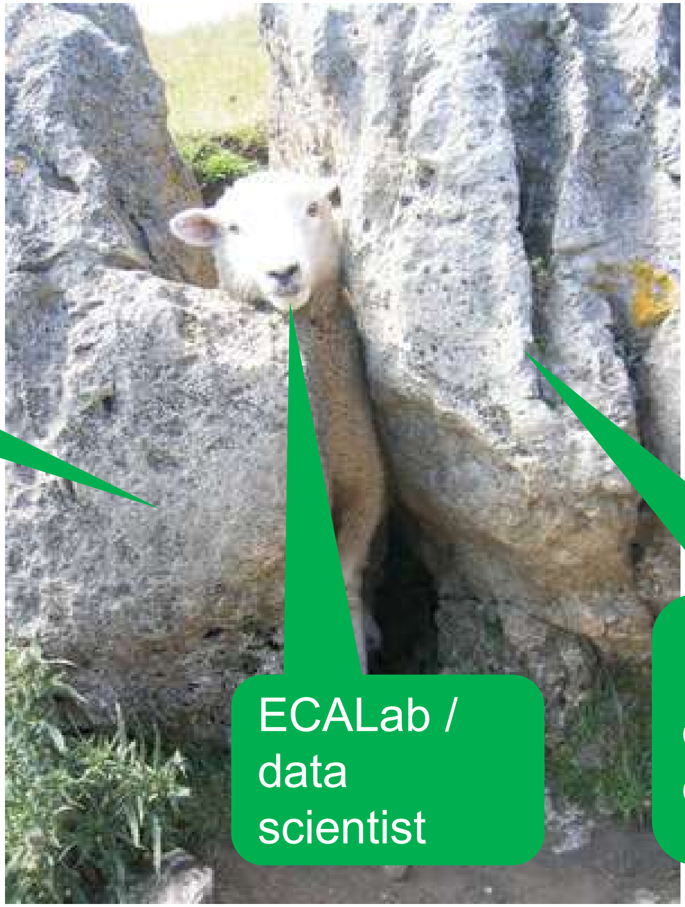


Business area / audit chambers

ECALab / data scientist

IT people and database experts

Small detour (between rock and a hard place)



Business area / audit chambers

ECALab / data scientist

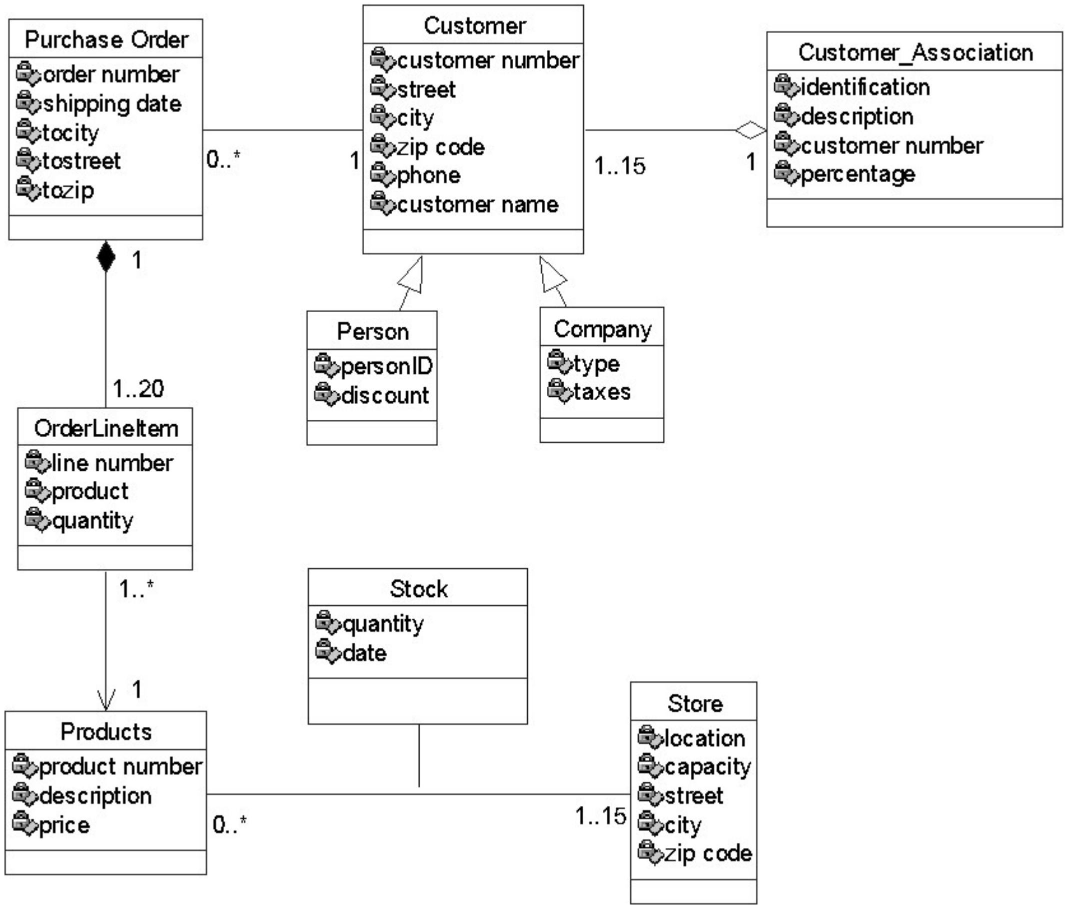
IT staff and database experts

Data protection officer

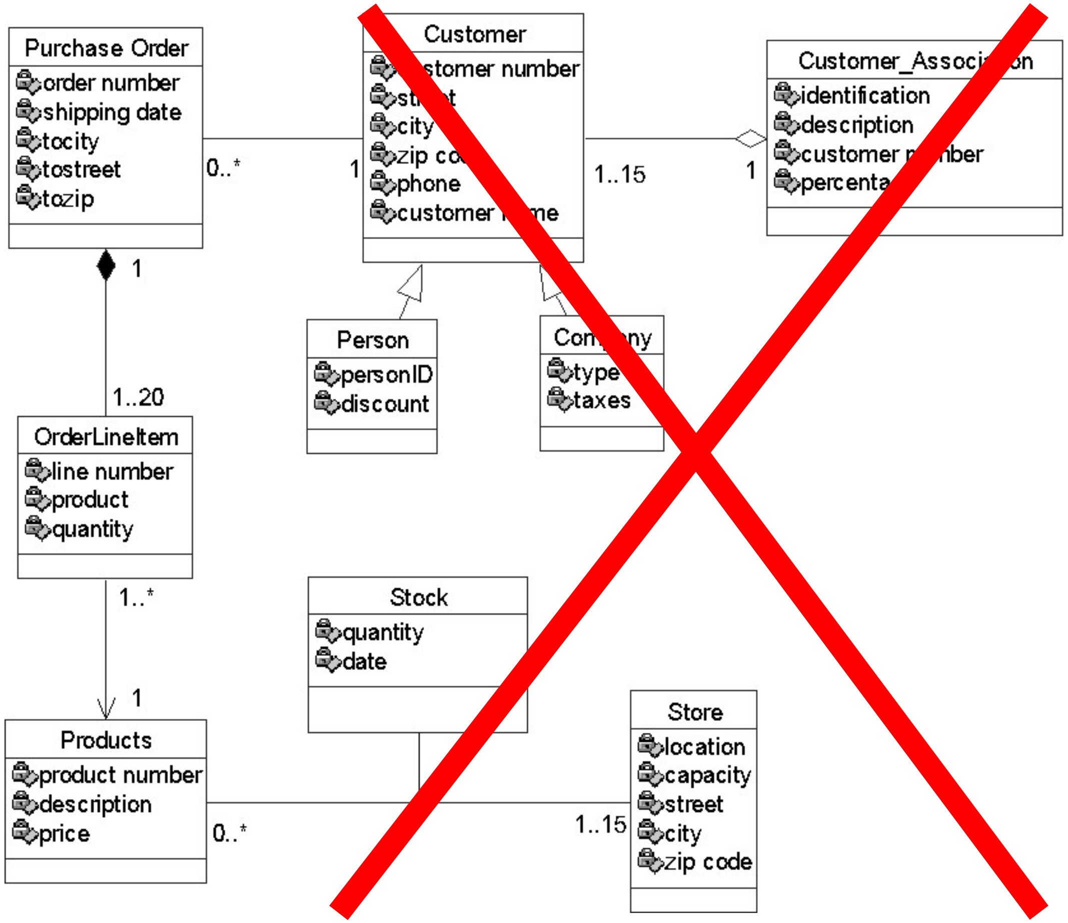
Access to data + data protection concerns

- Read-only access to the raw database of the Commission
    - Not much use without a data dictionary

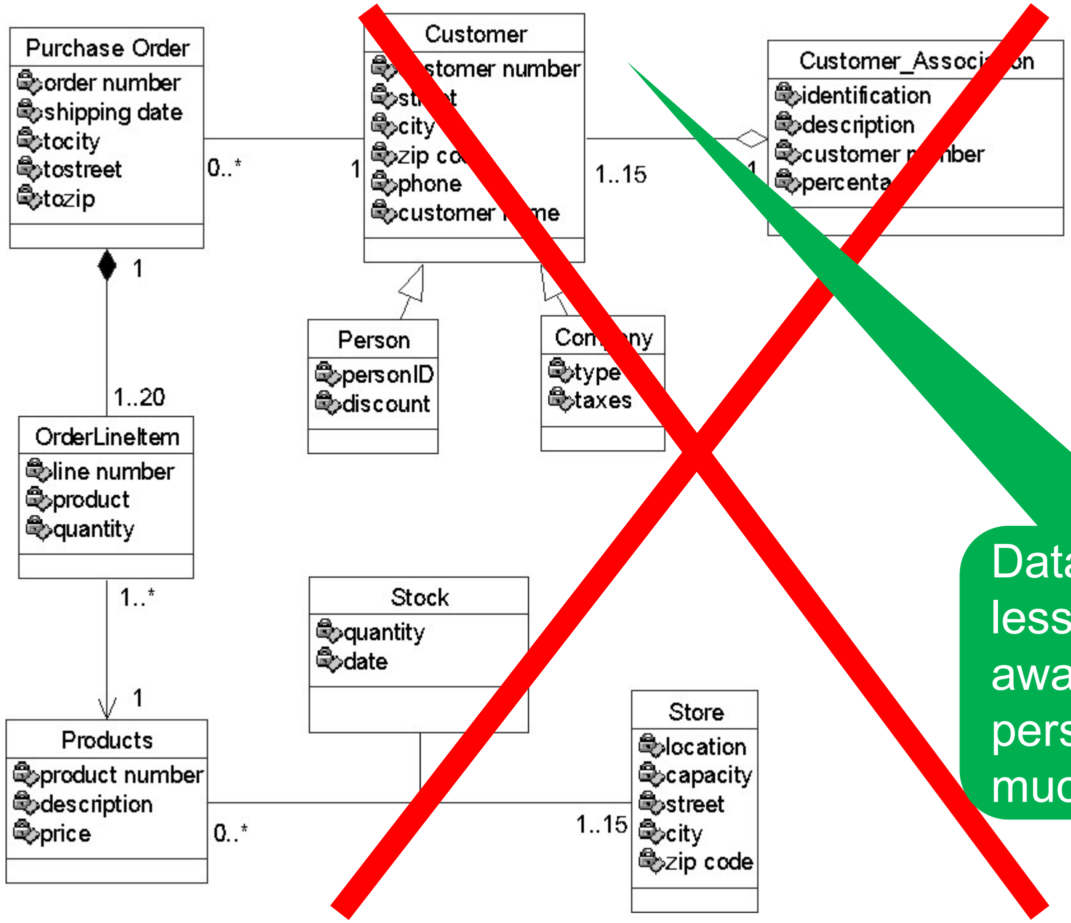
Relational database model



What you are likely to get for process mining

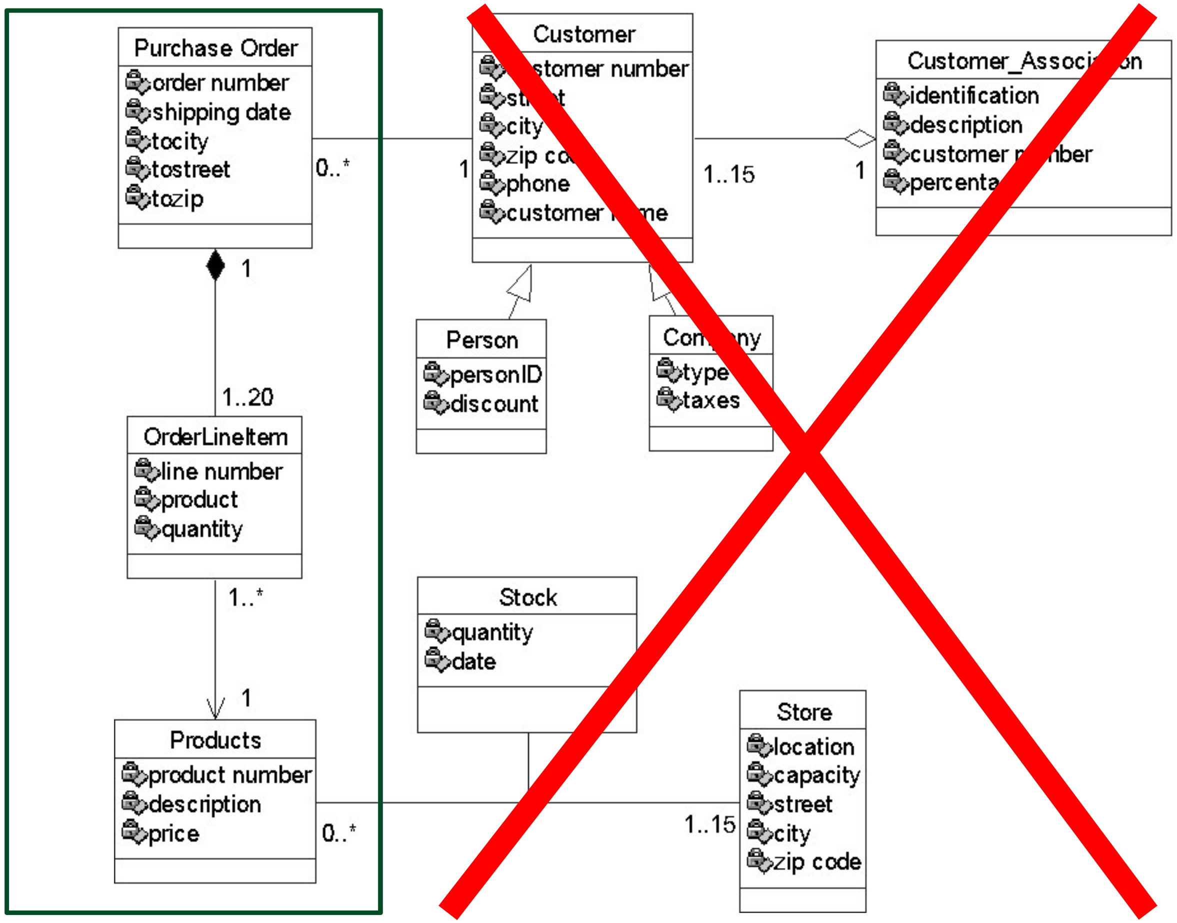


What you are likely to get for process mining

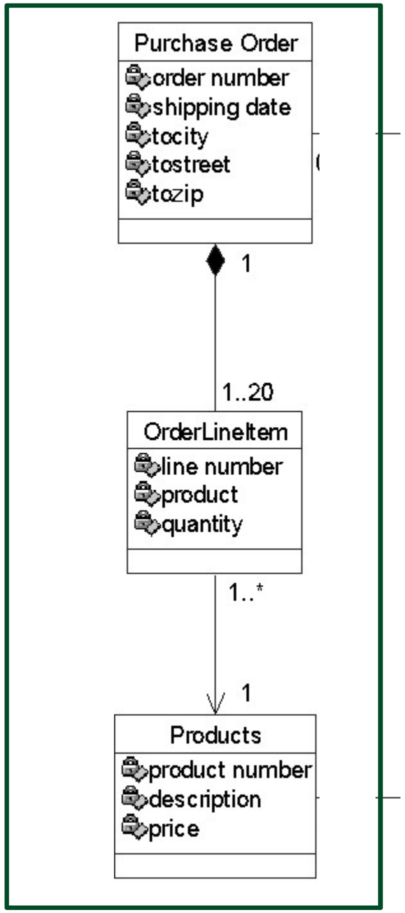


Data protection lesson 2: Stay away from personal data as much as possible!!

What you are likely to get for process mining



What you are likely to get for process mining



Access to data + data protection concerns

- Read-only access to the raw database of the Commission
    - Not much use without a data dictionary
- Access to data warehouse
    - DW reports (that I’ve seen) do not contain workflow information

Access to data + data protection concerns

- Read-only access to the raw database of the Commission
    - Not much use without a data dictionary
- Access to data warehouse
    - DW reports (that I’ve seen) do not contain workflow information
- Data may only be accessible through pdf exports
    - No meaningful database connection or custom legacy systems

Access to data + data protection concerns

- Read-only access to the raw database of the Commission
    - Not much use without a data dictionary
- Access to data warehouse
    - DW reports (that I've seen) do not contain workflow information
- Data may only be accessible through pdf exports
    - No meaningful database connection or custom legacy systems
- Dataset contains user IDs and agency names
    - We are allowed to process such information for audit purposes, but I cannot show you real data in my presentation
    - I created a synthetic dataset based on the original, fully anonymised, with fictional agency names, and added some fictional anomalies

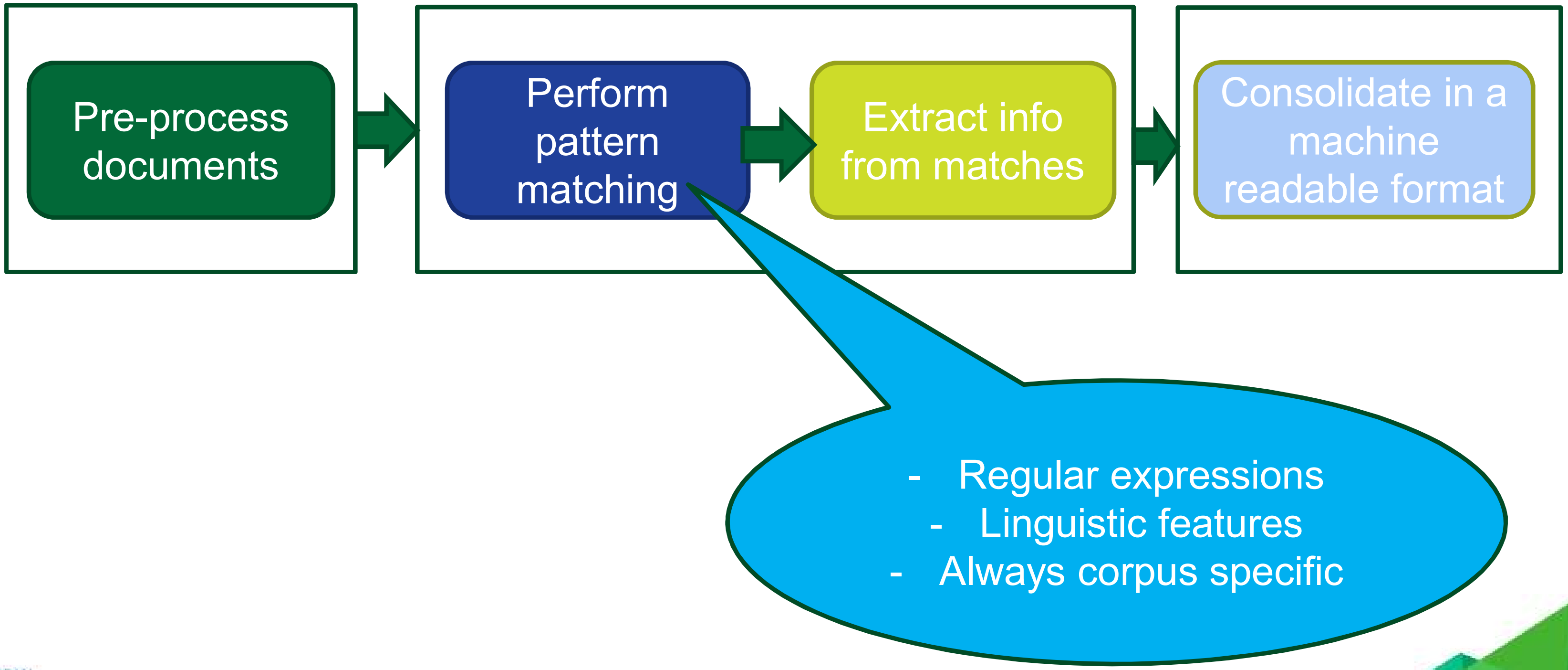
Audit of EU agencies

Information extraction from pdfs

Why?

- PDF is evil. Although it is called a PDF “document”, it’s nothing like Word or HTML document.
- PDF is more like a graphic representation. PDF contents are just a bunch of instructions that tell how to place the stuff at each exact position on a display or paper.
- In most cases, it has no logical structure such as sentences or paragraphs and it cannot even identify a text box properly.
- PDFs require a disproportionate amount of resources to extract data in a structured format.
- PDFs are not meant for electronic processing. They are meant for printing!

Information extraction pipeline



Pre-process documents

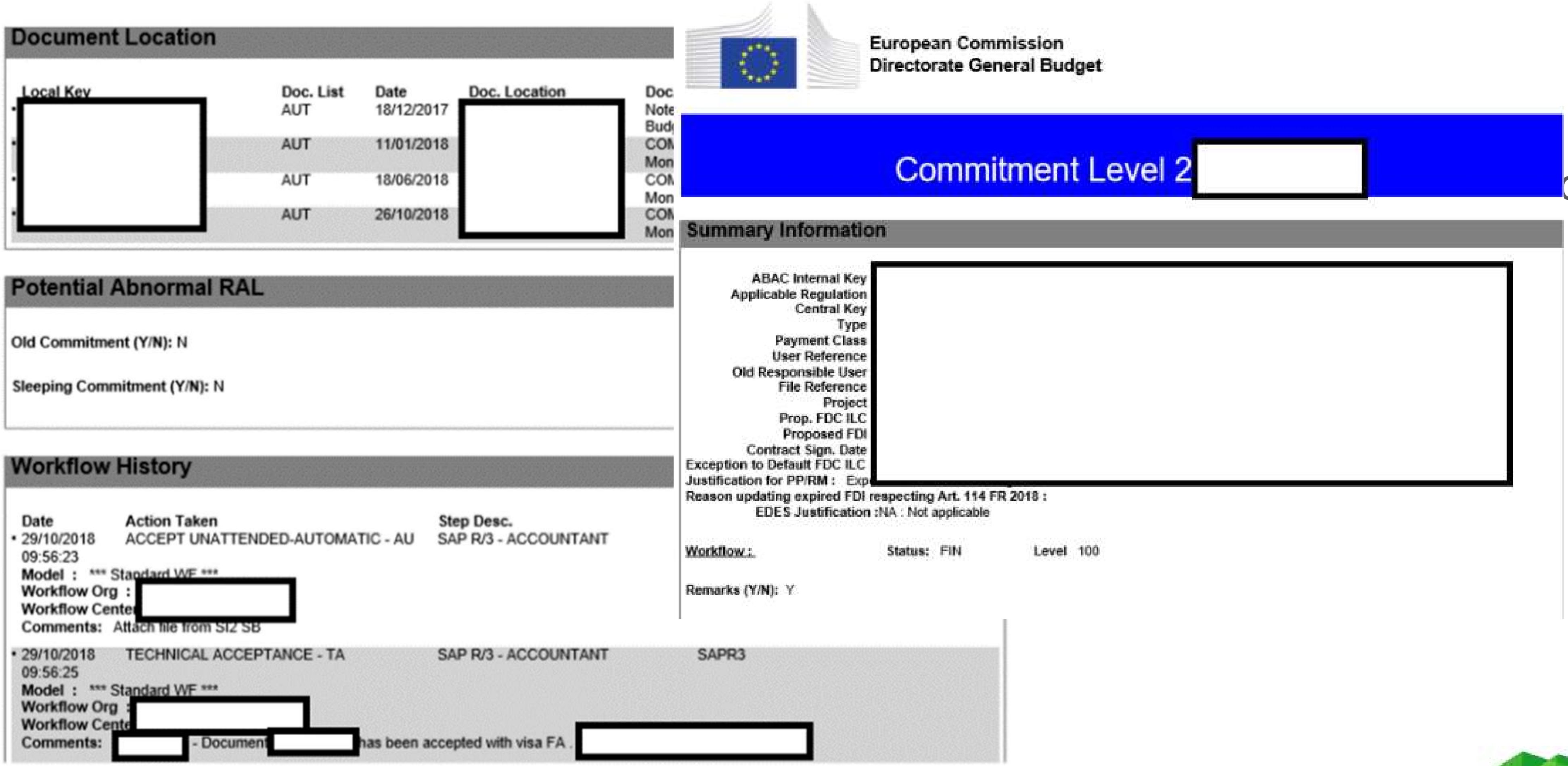
Perform pattern matching

Extract info from matches

Consolidate in a machine readable format

- Regular expressions
- Linguistic features
- Always corpus specific

Agency payment transactions



Document Location

Local Key	Doc. List	Date	Doc. Location	Doc
[redacted]	AUT	18/12/2017	[redacted]	Note Bud
[redacted]	AUT	11/01/2018	[redacted]	CO Mon
[redacted]	AUT	18/06/2018	[redacted]	CO Mon
[redacted]	AUT	26/10/2018	[redacted]	CO Mon

Potential Abnormal RAL

Old Commitment (Y/N): N

Sleeping Commitment (Y/N): N

Workflow History

Date	Action Taken	Step Desc.	
29/10/2018 09:56:23	ACCEPT UNATTENDED-AUTOMATIC - AU	SAP R/3 - ACCOUNTANT	
29/10/2018 09:56:25	TECHNICAL ACCEPTANCE - TA	SAP R/3 - ACCOUNTANT	SAPR3

- 29/10/2018 09:56:23 ACCEPT UNATTENDED-AUTOMATIC - AU SAP R/3 - ACCOUNTANT
  Model : *** Standard WF ***
  Workflow Org : [redacted]
  Workflow Cente
  Comments: Attach file from SI2 SB
- 29/10/2018 09:56:25 TECHNICAL ACCEPTANCE - TA SAP R/3 - ACCOUNTANT SAPR3
  Model : *** Standard WF ***
  Workflow Org [redacted]
  Workflow Cente
  Comments: [redacted] - Document [redacted] has been accepted with visa FA . [redacted]

European Commission
Directorate General Budget

Commitment Level 2 [redacted]

Summary Information

ABAC Internal Key
Applicable Regulation
Central Key
Type
Payment Class
User Reference
Old Responsible User
File Reference
Project
Prop. FDC ILC
Proposed FDI
Contract Sign. Date
Exception to Default FDC ILC
Justification for PP/RM : Exp [redacted]
Reason updating expired FDI respecting Art. 114 FR 2018 :
EDES Justification :NA : Not applicable

Workflow : **Status:** FIN **Level** 100

Remarks (Y/N): Y

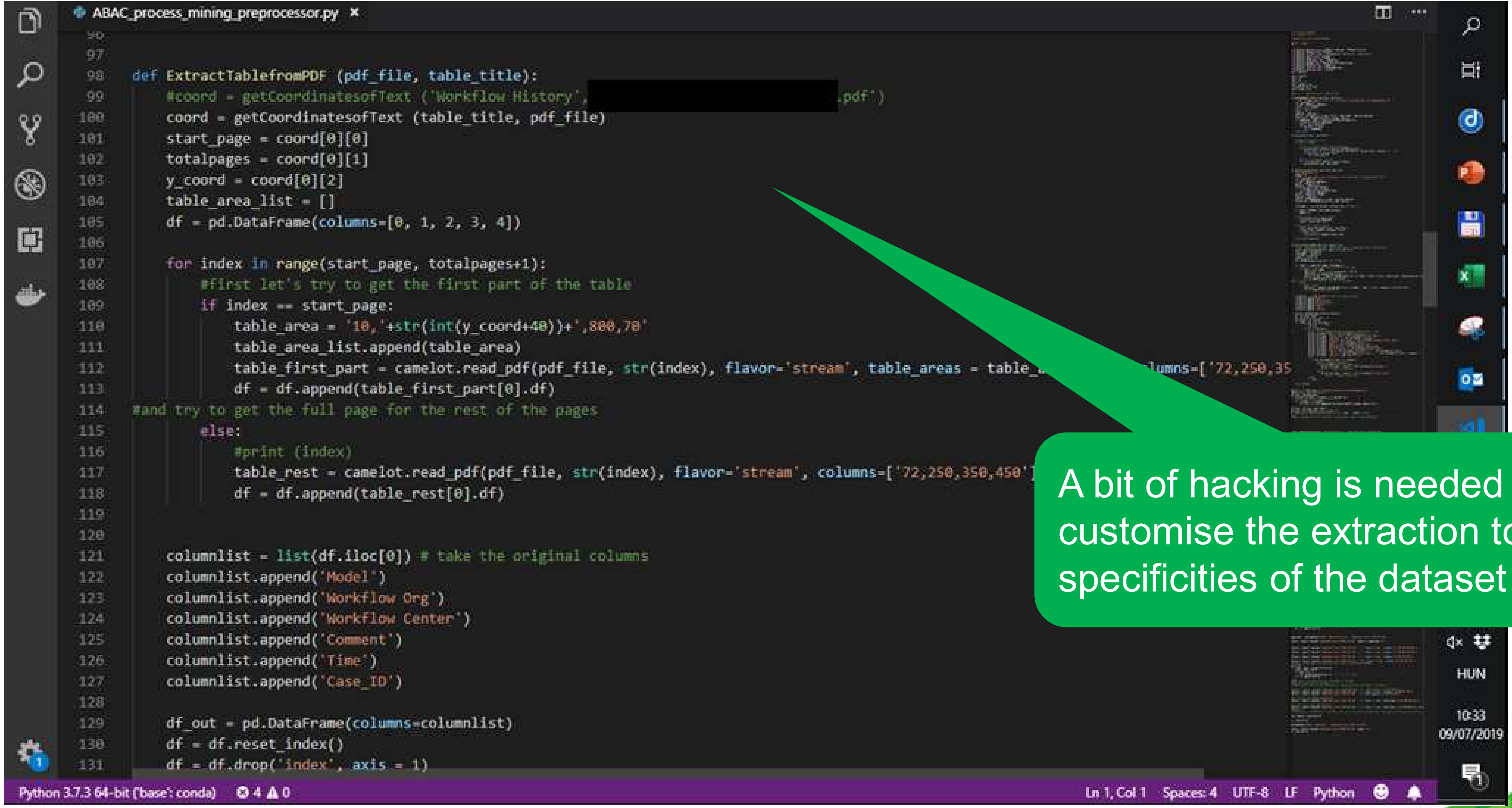
cal

A possible pdf data extraction workflow

- Camelot + pdfMiner libraries for Python
  - only works with text-based PDFs and not scanned documents
  - each table is a pandas DataFrame, which seamlessly integrates into ETL and data analysis workflows.

- My recommendation after much sweat and blood:
  - use pdfMiner to find and identify the part of the document containing the table you need
  - Use Camelot-py with the parameters you got from pdfMiner to extract the table
  - each dataset requires fine-tuning, exception handling and post-processing

Python code for data extraction


ABAC_process_mining_preprocessor.py
Python 3.7.3 64-bit ('base': conda)
⊗ 4 ⚠ 0
Ln 1, Col 1
Spaces: 4
UTF-8
LF
Python
HUN
10:33
09/07/2019

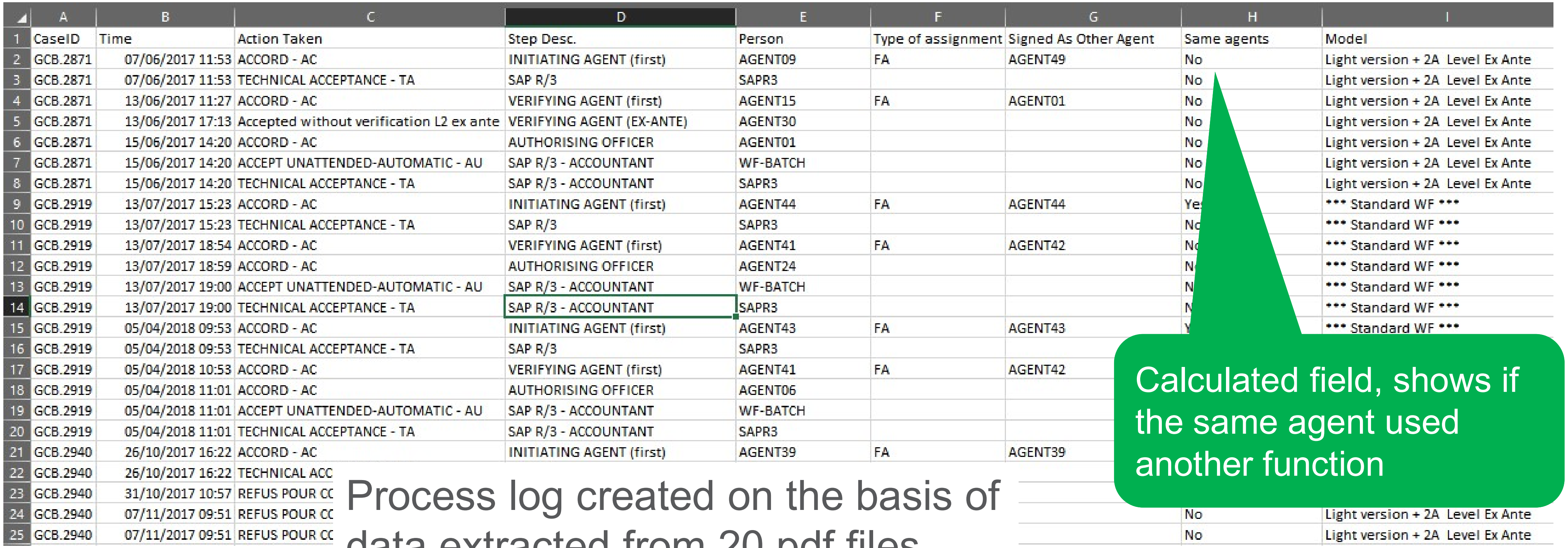
```python
def ExtractTablefromPDF (pdf_file, table_title):
    #coord = getCoordinatesofText ('Workflow History', [redacted].pdf')
    coord = getCoordinatesofText (table_title, pdf_file)
    start_page = coord[0][0]
    totalpages = coord[0][1]
    y_coord = coord[0][2]
    table_area_list = []
    df = pd.DataFrame(columns=[0, 1, 2, 3, 4])

    for index in range(start_page, totalpages+1):
        #first let's try to get the first part of the table
        if index == start_page:
            table_area = '10,'+str(int(y_coord+40))+',800,70'
            table_area_list.append(table_area)
            table_first_part = camelot.read_pdf(pdf_file, str(index), flavor='stream', table_areas = table_          lumns=['72,250,35
            df = df.append(table_first_part[0].df)
#and try to get the full page for the rest of the pages
        else:
            #print (index)
            table_rest = camelot.read_pdf(pdf_file, str(index), flavor='stream', columns=['72,250,350,450']
            df = df.append(table_rest[0].df)


    columnlist = list(df.iloc[0]) # take the original columns
    columnlist.append('Model')
    columnlist.append('Workflow Org')
    columnlist.append('Workflow Center')
    columnlist.append('Comment')
    columnlist.append('Time')
    columnlist.append('Case_ID')

    df_out = pd.DataFrame(columns=columnlist)
    df = df.reset_index()
    df = df.drop('index', axis = 1)
```

A bit of hacking is needed to customise the extraction to the specificities of the dataset

DISCLAIMER:

- The data, all agency names, anomalies and incidents portrayed in this dataset and process mining experiment are fictitious. No identification with actual persons (living or deceased), positions, organisations and events is intended or should be inferred.
- Even though the structure of the data closely follows that of the European Commission’s accounting system, this presentation does not imply that the EC’s control systems allow for any of the shown anomalies to occur.

ABAC workflow data



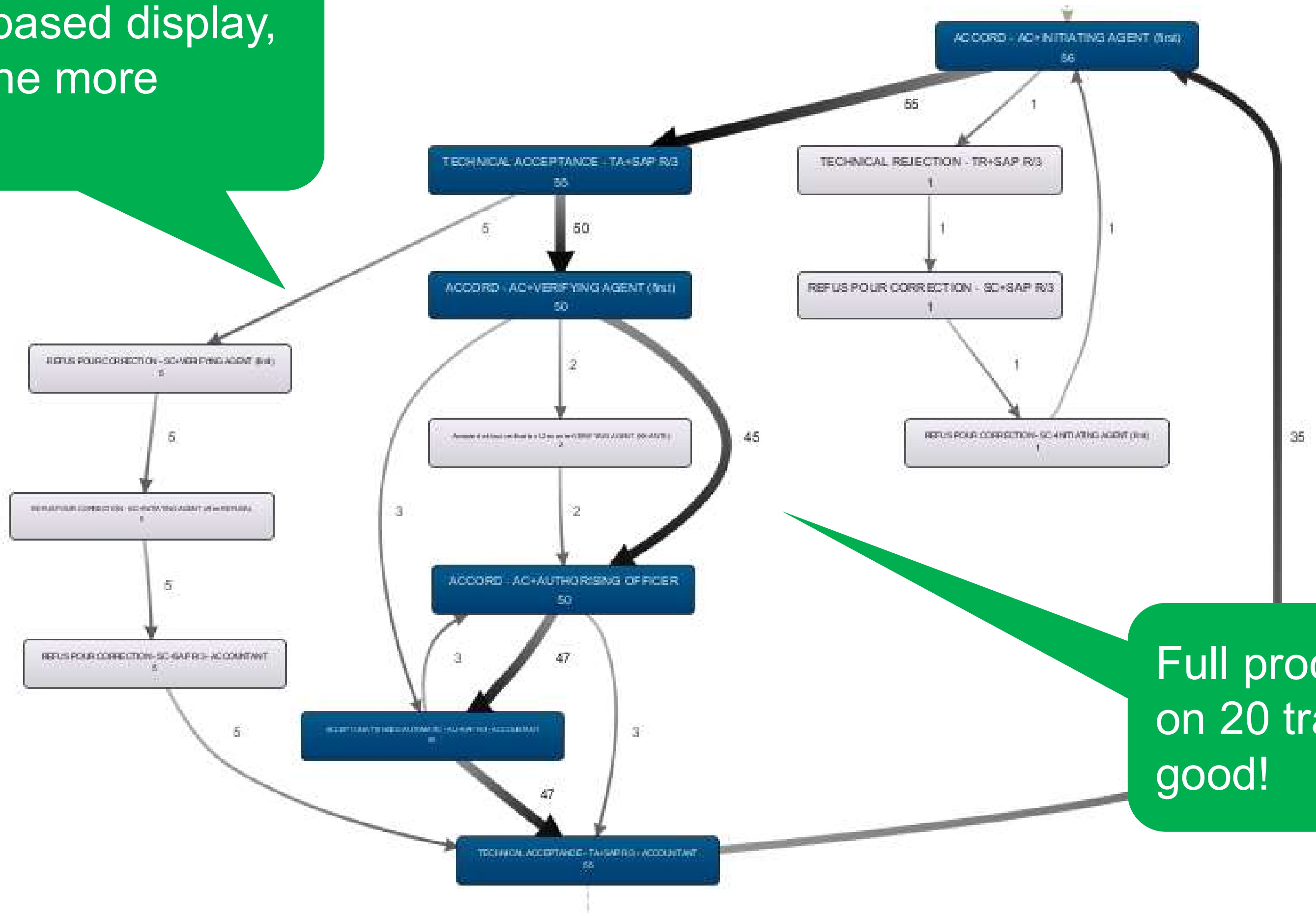
	A	B	C	D	E	F	G	H	I
1	CaseID	Time	Action Taken	Step Desc.	Person	Type of assignment	Signed As Other Agent	Same agents	Model
2	GCB.2871	07/06/2017 11:53	ACCORD - AC	INITIATING AGENT (first)	AGENT09	FA	AGENT49	No	Light version + 2A Level Ex Ante
3	GCB.2871	07/06/2017 11:53	TECHNICAL ACCEPTANCE - TA	SAP R/3	SAPR3			No	Light version + 2A Level Ex Ante
4	GCB.2871	13/06/2017 11:27	ACCORD - AC	VERIFYING AGENT (first)	AGENT15	FA	AGENT01	No	Light version + 2A Level Ex Ante
5	GCB.2871	13/06/2017 17:13	Accepted without verification L2 ex ante	VERIFYING AGENT (EX-ANTE)	AGENT30			No	Light version + 2A Level Ex Ante
6	GCB.2871	15/06/2017 14:20	ACCORD - AC	AUTHORISING OFFICER	AGENT01			No	Light version + 2A Level Ex Ante
7	GCB.2871	15/06/2017 14:20	ACCEPT UNATTENDED-AUTOMATIC - AU	SAP R/3 - ACCOUNTANT	WF-BATCH			No	Light version + 2A Level Ex Ante
8	GCB.2871	15/06/2017 14:20	TECHNICAL ACCEPTANCE - TA	SAP R/3 - ACCOUNTANT	SAPR3			No	Light version + 2A Level Ex Ante
9	GCB.2919	13/07/2017 15:23	ACCORD - AC	INITIATING AGENT (first)	AGENT44	FA	AGENT44	Ye	*** Standard WF ***
10	GCB.2919	13/07/2017 15:23	TECHNICAL ACCEPTANCE - TA	SAP R/3	SAPR3			Nc	*** Standard WF ***
11	GCB.2919	13/07/2017 18:54	ACCORD - AC	VERIFYING AGENT (first)	AGENT41	FA	AGENT42	Nc	*** Standard WF ***
12	GCB.2919	13/07/2017 18:59	ACCORD - AC	AUTHORISING OFFICER	AGENT24			N	*** Standard WF ***
13	GCB.2919	13/07/2017 19:00	ACCEPT UNATTENDED-AUTOMATIC - AU	SAP R/3 - ACCOUNTANT	WF-BATCH			N	*** Standard WF ***
14	GCB.2919	13/07/2017 19:00	TECHNICAL ACCEPTANCE - TA	SAP R/3 - ACCOUNTANT	SAPR3			N	*** Standard WF ***
15	GCB.2919	05/04/2018 09:53	ACCORD - AC	INITIATING AGENT (first)	AGENT43	FA	AGENT43	Y	*** Standard WF ***
16	GCB.2919	05/04/2018 09:53	TECHNICAL ACCEPTANCE - TA	SAP R/3	SAPR3				
17	GCB.2919	05/04/2018 10:53	ACCORD - AC	VERIFYING AGENT (first)	AGENT41	FA	AGENT42		
18	GCB.2919	05/04/2018 11:01	ACCORD - AC	AUTHORISING OFFICER	AGENT06				
19	GCB.2919	05/04/2018 11:01	ACCEPT UNATTENDED-AUTOMATIC - AU	SAP R/3 - ACCOUNTANT	WF-BATCH				
20	GCB.2919	05/04/2018 11:01	TECHNICAL ACCEPTANCE - TA	SAP R/3 - ACCOUNTANT	SAPR3				
21	GCB.2940	26/10/2017 16:22	ACCORD - AC	INITIATING AGENT (first)	AGENT39	FA	AGENT39		
22	GCB.2940	26/10/2017 16:22	TECHNICAL ACC						
23	GCB.2940	31/10/2017 10:57	REFUS POUR CC						
24	GCB.2940	07/11/2017 09:51	REFUS POUR CC					No	Light version + 2A Level Ex Ante
25	GCB.2940	07/11/2017 09:51	REFUS POUR CC					No	Light version + 2A Level Ex Ante

Calculated field, shows if the same agent used another function

Process log created on the basis of data extracted from 20 pdf files, semi-randomly chosen by auditors among various agencies

ABAC workflow data

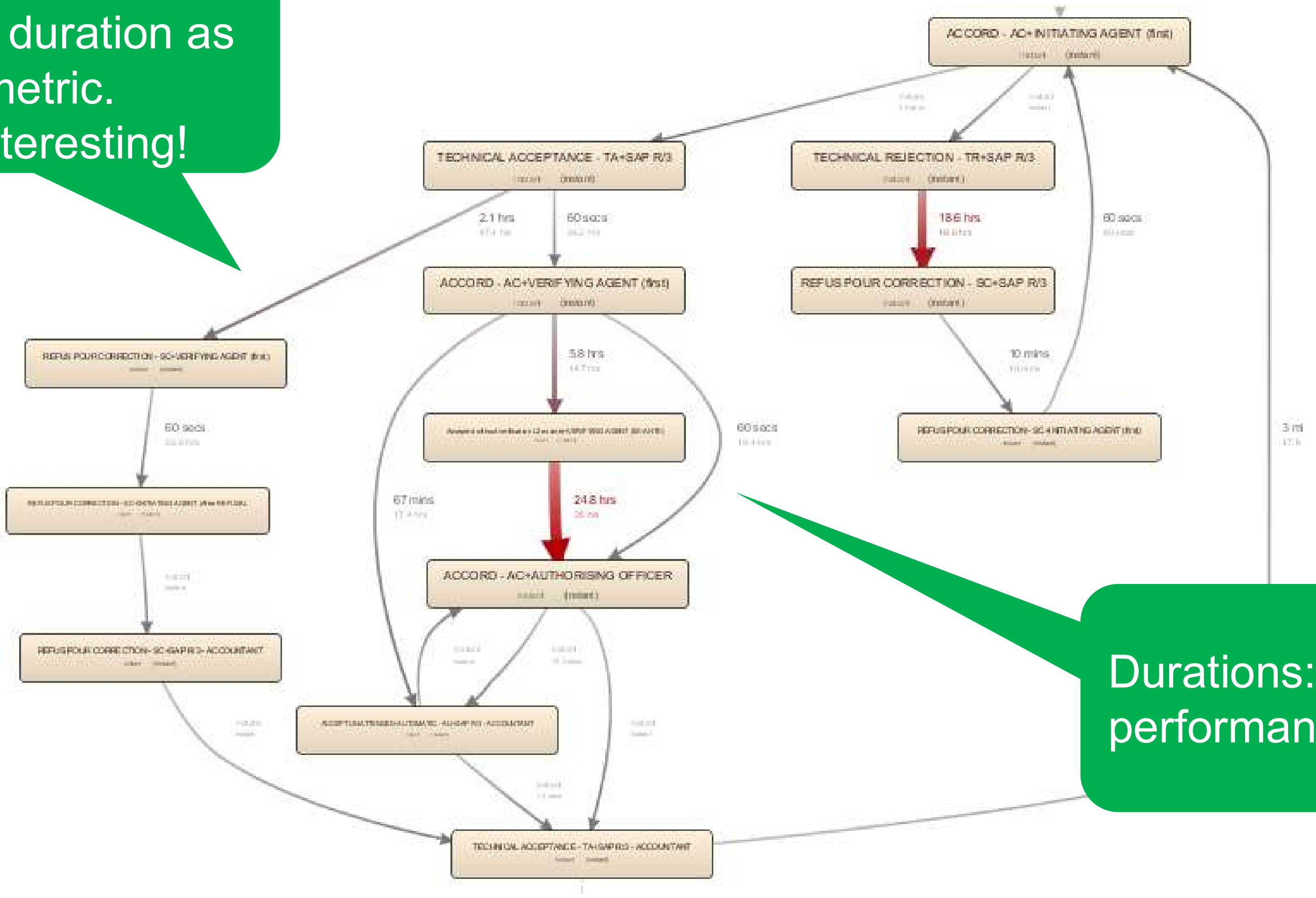
Frequency based display, the darker the more frequent


ACCORD - AC+INITIATING AGENT (first)
56
55
1
TECHNICAL ACCEPTANCE - TA+SAP R/3
55
TECHNICAL REJECTION - TR+SAP R/3
1
5
50
1
1
ACCORD - AC+VERIFYING AGENT (first)
50
REFUS POUR CORRECTION - SC+SAP R/3
1
REFUS POUR CORRECTION - SC+VERIFYING AGENT (first)
5
2
1
5
45
REFUS POUR CORRECTION - SC+INITIATING AGENT (first)
1
35
3
2
ACCORD - AC+AUTHORISING OFFICER
50
5
REFUS POUR CORRECTION - SC+SAP R/3 - ACCOUNTANT
5
3
47
5
3
47
TECHNICAL ACCEPTANCE - TA+SAP R/3 - ACCOUNTANT
55

Full process model based on 20 transactions. Looks good!

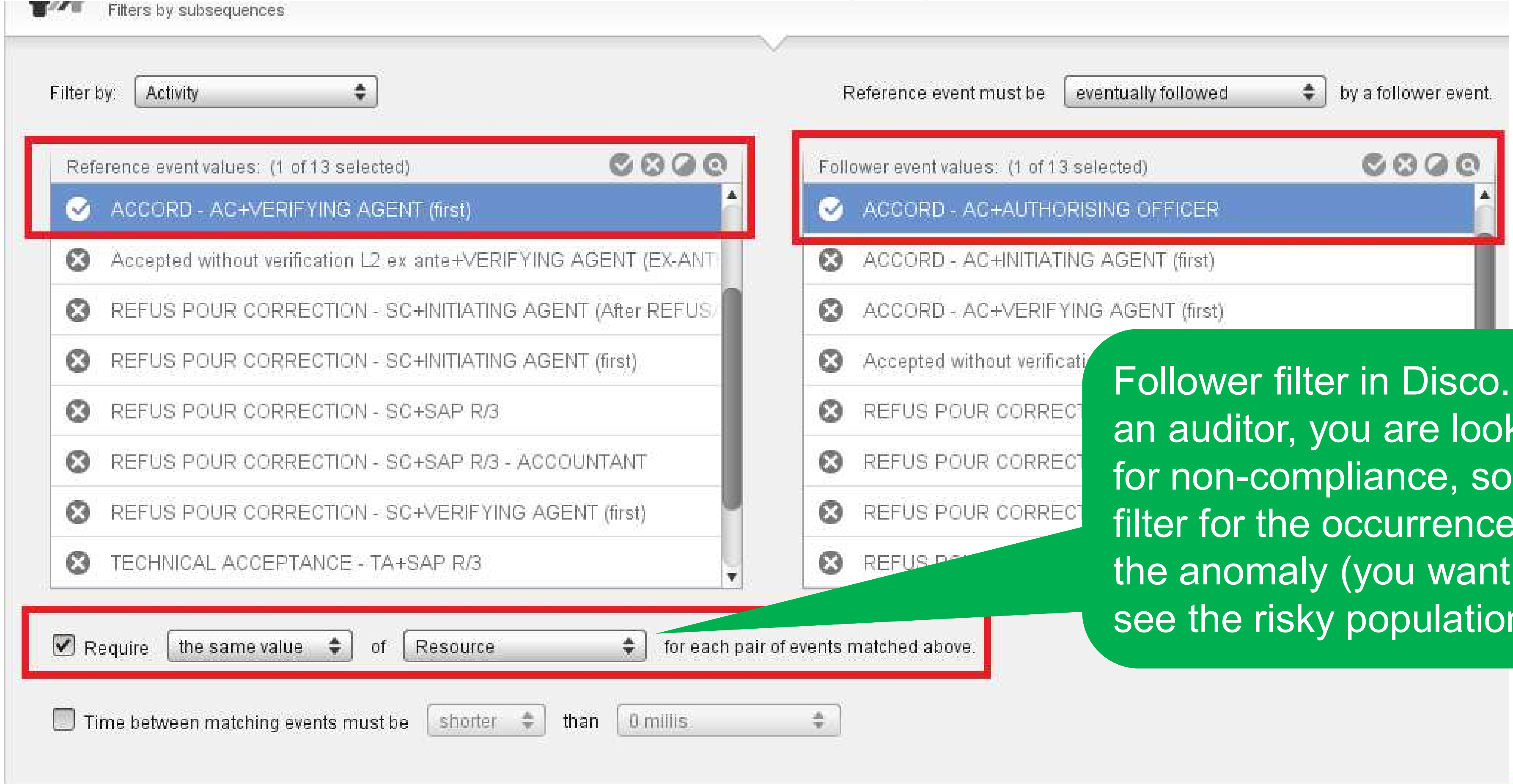
ABAC workflow data

Displaying mean duration, with minimum duration as a secondary metric. Outliers are interesting!


ACCORD - AC+INITIATING AGENT (first)
TECHNICAL ACCEPTANCE - TA+SAP R/3
TECHNICAL REJECTION - TR+SAP R/3
2.1 hrs
60 secs
18.6 hrs
60 secs
ACCORD - AC+VERIFYING AGENT (first)
REFUS POUR CORRECTION - SC+SAP R/3
REFUS POUR CORRECTION - SC+VERIFYING AGENT (first)
5.8 hrs
10 mins
60 secs
60 secs
REFUS POUR CORRECTION - SC+INITIATING AGENT (first)
3 mi
67 mins
24.8 hrs
ACCORD - AC+AUTHORISING OFFICER
REFUS POUR CORRECTION - SC+SAP R/3 - ACCOUNTANT
TECHNICAL ACCEPTANCE - TA+SAP R/3 - ACCOUNTANT

Durations: first step of performance checking

Compliance check: segregation of duties



Filters by subsequences

Filter by: Activity

Reference event must be eventually followed by a follower event.

Reference event values: (1 of 13 selected)

- ✓ ACCORD - AC+VERIFYING AGENT (first)
- ⊗ Accepted without verification L2 ex ante+VERIFYING AGENT (EX-ANT
- ⊗ REFUS POUR CORRECTION - SC+INITIATING AGENT (After REFUS
- ⊗ REFUS POUR CORRECTION - SC+INITIATING AGENT (first)
- ⊗ REFUS POUR CORRECTION - SC+SAP R/3
- ⊗ REFUS POUR CORRECTION - SC+SAP R/3 - ACCOUNTANT
- ⊗ REFUS POUR CORRECTION - SC+VERIFYING AGENT (first)
- ⊗ TECHNICAL ACCEPTANCE - TA+SAP R/3

Follower event values: (1 of 13 selected)

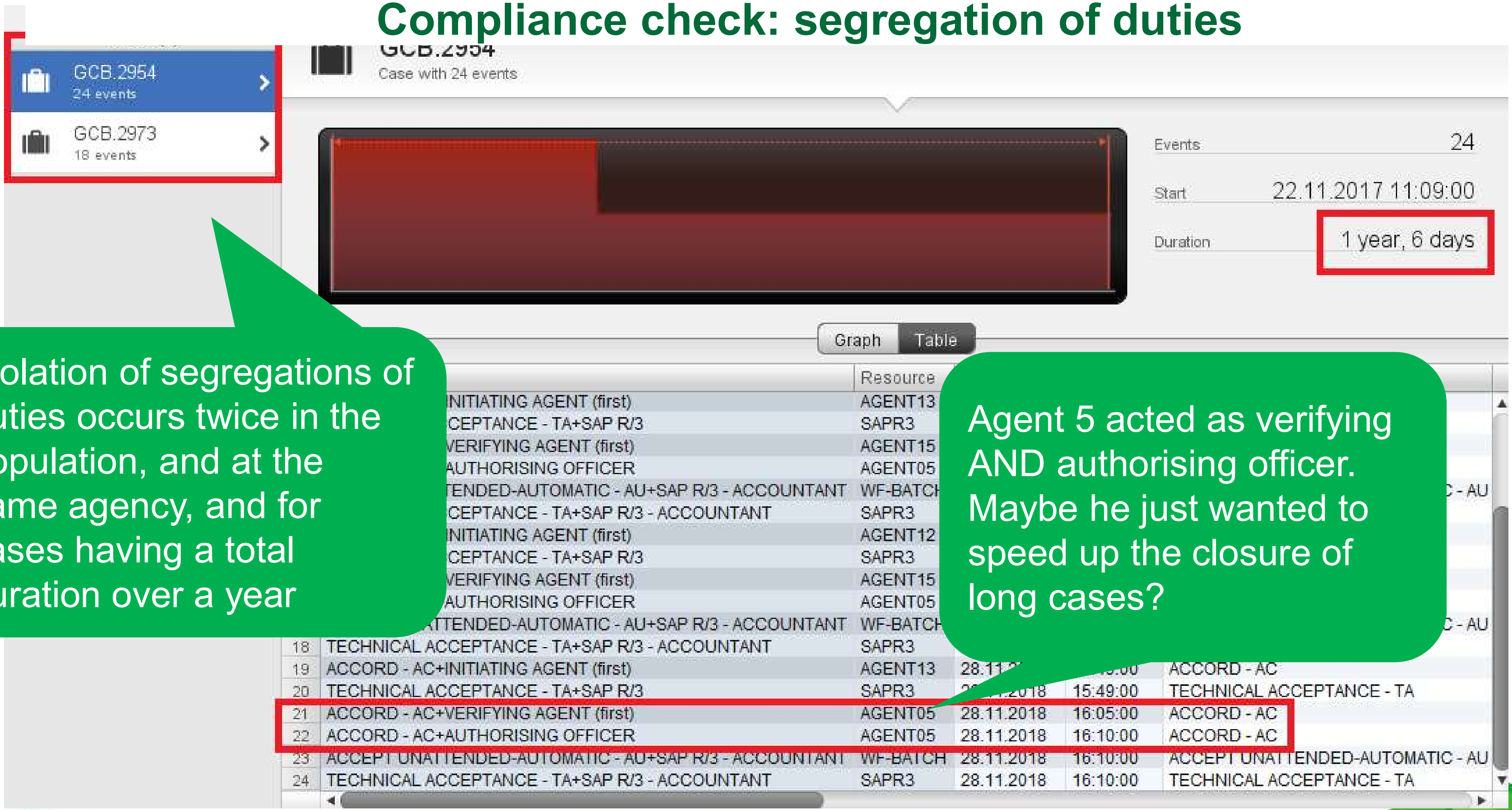
- ✓ ACCORD - AC+AUTHORISING OFFICER
- ⊗ ACCORD - AC+INITIATING AGENT (first)
- ⊗ ACCORD - AC+VERIFYING AGENT (first)
- ⊗ Accepted without verificati
- ⊗ REFUS POUR CORREC
- ⊗ REFUS POUR CORREC
- ⊗ REFUS POUR CORREC
- ⊗ REFUS P

☑ Require the same value of Resource for each pair of events matched above.

☐ Time between matching events must be shorter than 0 millis

Follower filter in Disco. As an auditor, you are looking for non-compliance, so you filter for the occurrence of the anomaly (you want to see the risky population)

Compliance check: segregation of duties



GCB.2954
Case with 24 events

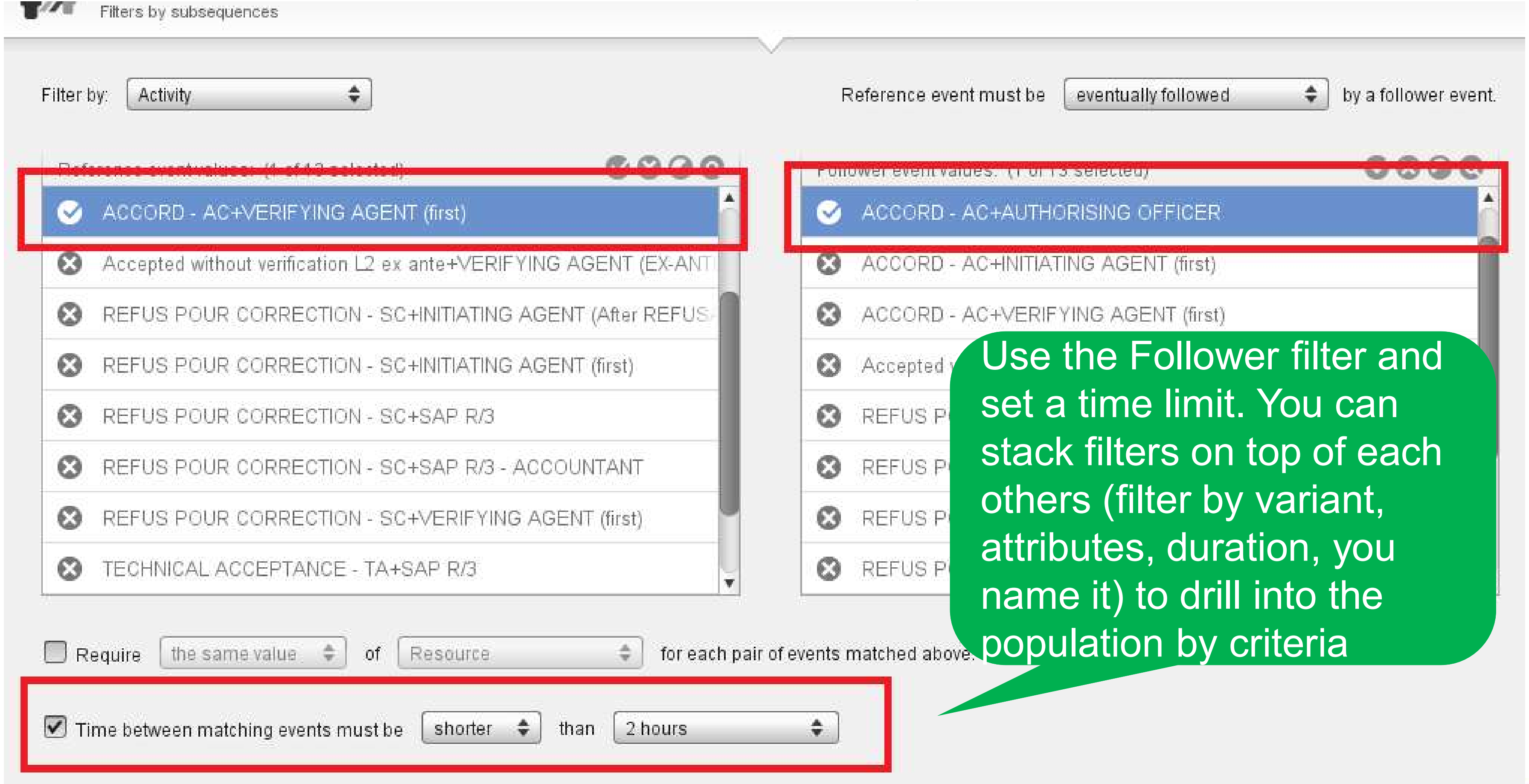
Events 24
Start 22.11.2017 11:09:00
Duration 1 year, 6 days

		Resource			
	INITIATING AGENT (first)	AGENT13			
	CEPTANCE - TA+SAP R/3	SAPR3			
	VERIFYING AGENT (first)	AGENT15			
	AUTHORISING OFFICER	AGENT05			
	TENDED-AUTOMATIC - AU+SAP R/3 - ACCOUNTANT	WF-BATC			C - AU
	CEPTANCE - TA+SAP R/3 - ACCOUNTANT	SAPR3			
	INITIATING AGENT (first)	AGENT12			
	CEPTANCE - TA+SAP R/3	SAPR3			
	VERIFYING AGENT (first)	AGENT15			
	AUTHORISING OFFICER	AGENT05			
	TTENDED-AUTOMATIC - AU+SAP R/3 - ACCOUNTANT	WF-BATC			C - AU
18	TECHNICAL ACCEPTANCE - TA+SAP R/3 - ACCOUNTANT	SAPR3			
19	ACCORD - AC+INITIATING AGENT (first)	AGENT13	28.11		ACCORD - AC
20	TECHNICAL ACCEPTANCE - TA+SAP R/3	SAPR3	.2018	15:49:00	TECHNICAL ACCEPTANCE - TA
21	ACCORD - AC+VERIFYING AGENT (first)	AGENT05	28.11.2018	16:05:00	ACCORD - AC
22	ACCORD - AC+AUTHORISING OFFICER	AGENT05	28.11.2018	16:10:00	ACCORD - AC
23	ACCEPT UNATTENDED-AUTOMATIC - AU+SAP R/3 - ACCOUNTANT	WF-BATCH	28.11.2018	16:10:00	ACCEPT UNATTENDED-AUTOMATIC - AU
24	TECHNICAL ACCEPTANCE - TA+SAP R/3 - ACCOUNTANT	SAPR3	28.11.2018	16:10:00	TECHNICAL ACCEPTANCE - TA

Violation of segregations of duties occurs twice in the population, and at the same agency, and for cases having a total duration over a year

Agent 5 acted as verifying AND authorising officer. Maybe he just wanted to speed up the closure of long cases?

Compliance check: suspiciously short approval times



Filters by subsequences

Filter by: Activity

Reference event must be eventually followed by a follower event.

Reference event values: (1 of 13 selected)

- ✓ ACCORD - AC+VERIFYING AGENT (first)
- ✗ Accepted without verification L2 ex ante+VERIFYING AGENT (EX-ANT
- ✗ REFUS POUR CORRECTION - SC+INITIATING AGENT (After REFUS
- ✗ REFUS POUR CORRECTION - SC+INITIATING AGENT (first)
- ✗ REFUS POUR CORRECTION - SC+SAP R/3
- ✗ REFUS POUR CORRECTION - SC+SAP R/3 - ACCOUNTANT
- ✗ REFUS POUR CORRECTION - SC+VERIFYING AGENT (first)
- ✗ TECHNICAL ACCEPTANCE - TA+SAP R/3

Follower event values: (1 of 13 selected)

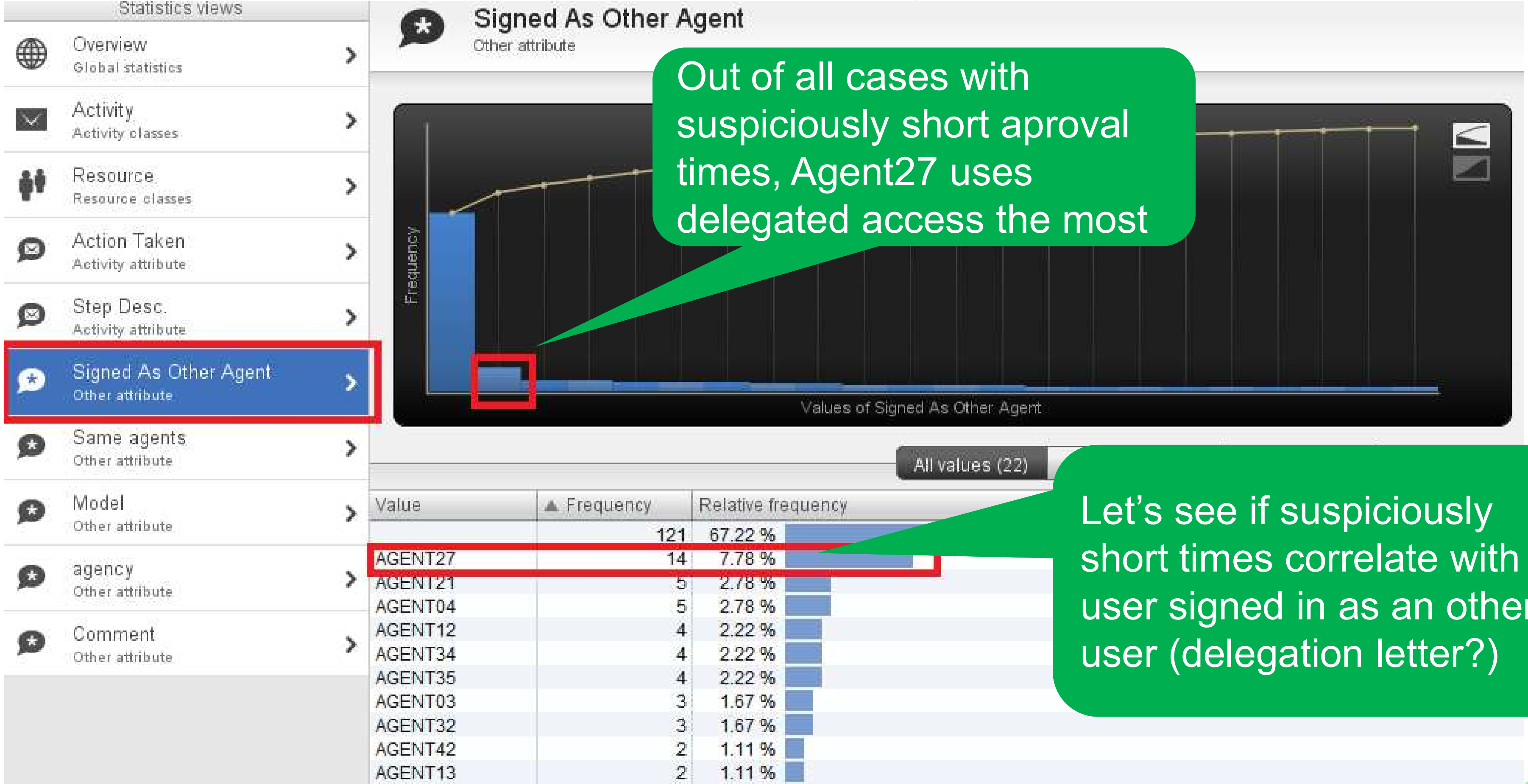
- ✓ ACCORD - AC+AUTHORISING OFFICER
- ✗ ACCORD - AC+INITIATING AGENT (first)
- ✗ ACCORD - AC+VERIFYING AGENT (first)
- ✗ Accepted
- ✗ REFUS P
- ✗ REFUS P
- ✗ REFUS P
- ✗ REFUS P

☐ Require the same value of Resource for each pair of events matched above.

☑ Time between matching events must be shorter than 2 hours

Use the Follower filter and set a time limit. You can stack filters on top of each others (filter by variant, attributes, duration, you name it) to drill into the population by criteria

Compliance check: suspiciously short approval times

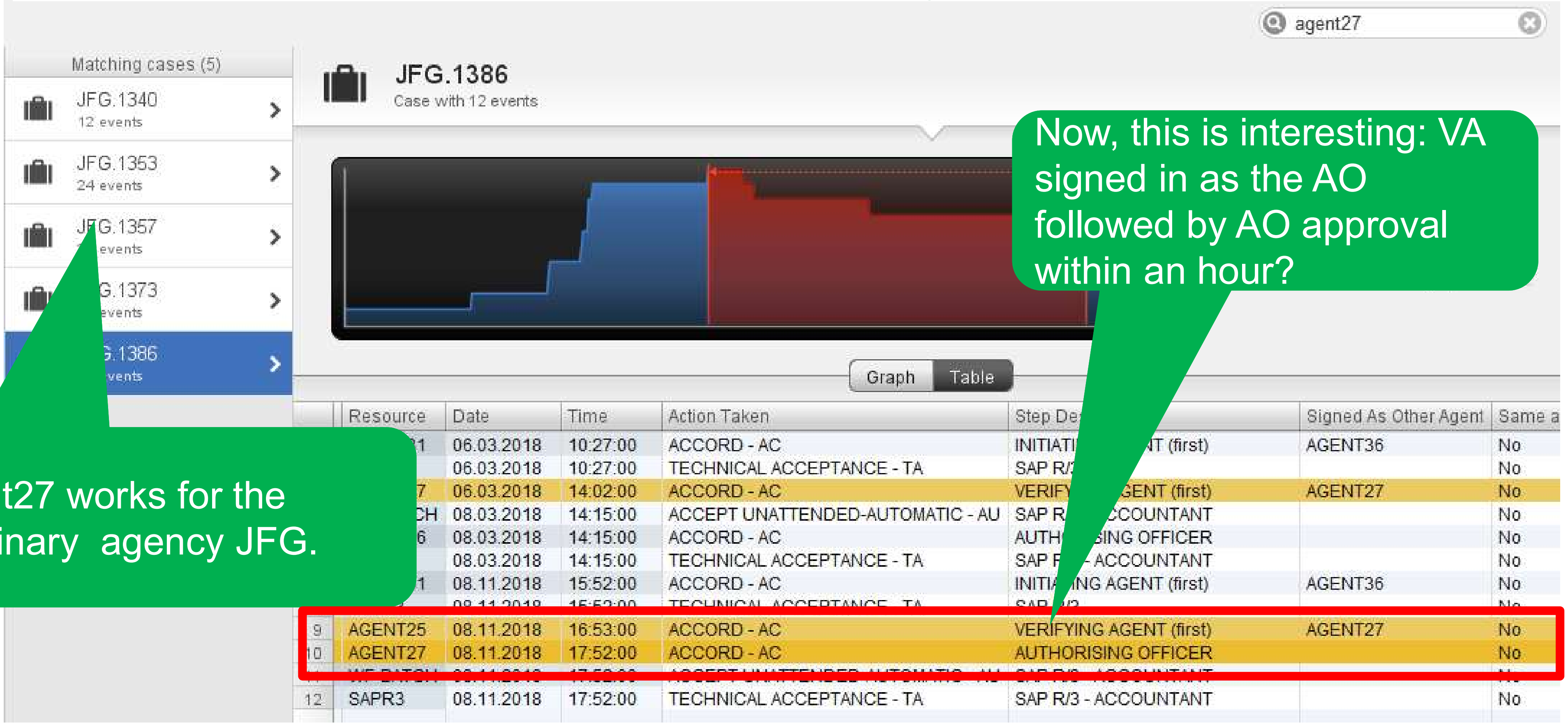


Value	▲ Frequency	Relative frequency
	121	67.22 %
AGENT27	14	7.78 %
AGENT21	5	2.78 %
AGENT04	5	2.78 %
AGENT12	4	2.22 %
AGENT34	4	2.22 %
AGENT35	4	2.22 %
AGENT03	3	1.67 %
AGENT32	3	1.67 %
AGENT42	2	1.11 %
AGENT13	2	1.11 %

Out of all cases with suspiciously short aproval times, Agent27 uses delegated access the most

Let’s see if suspiciously short times correlate with a user signed in as an other user (delegation letter?)

Compliance check: suspiciously short approval times

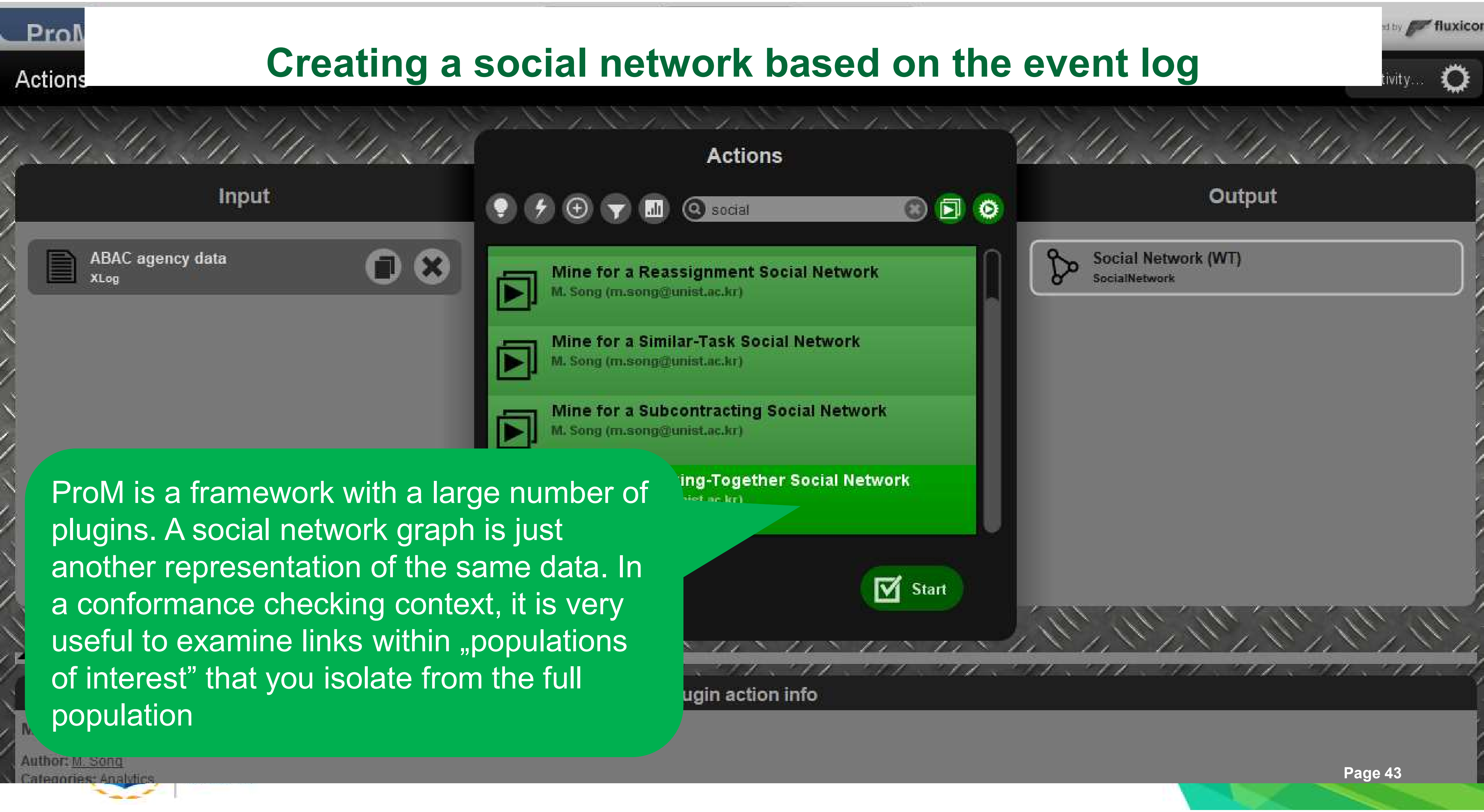


	Resource	Date	Time	Action Taken	Step De	Signed As Other Agent	Same a
	1	06.03.2018	10:27:00	ACCORD - AC	INITIATI T (first)	AGENT36	No
		06.03.2018	10:27:00	TECHNICAL ACCEPTANCE - TA	SAP R/		No
	7	06.03.2018	14:02:00	ACCORD - AC	VERIFY GENT (first)	AGENT27	No
	CH	08.03.2018	14:15:00	ACCEPT UNATTENDED-AUTOMATIC - AU	SAP R CCOUNTANT		No
	6	08.03.2018	14:15:00	ACCORD - AC	AUTH SING OFFICER		No
		08.03.2018	14:15:00	TECHNICAL ACCEPTANCE - TA	SAP R - ACCOUNTANT		No
	1	08.11.2018	15:52:00	ACCORD - AC	INITIA ING AGENT (first)	AGENT36	No
		08.11.2018	15:52:00	TECHNICAL ACCEPTANCE - TA	SAP R/3		No
9	AGENT25	08.11.2018	16:53:00	ACCORD - AC	VERIFYING AGENT (first)	AGENT27	No
10	AGENT27	08.11.2018	17:52:00	ACCORD - AC	AUTHORISING OFFICER		No
11	WF-BATCH	08.11.2018	17:52:00	ACCEPT UNATTENDED-AUTOMATIC - AU	SAP R/3 - ACCOUNTANT		No
12	SAPR3	08.11.2018	17:52:00	TECHNICAL ACCEPTANCE - TA	SAP R/3 - ACCOUNTANT		No

Now, this is interesting: VA signed in as the AO followed by AO approval within an hour?

Agent27 works for the imaginary agency JFG.

Creating a social network based on the event log



ABAC agency data
XLog

Mine for a Reassignment Social Network
M. Song (m.song@unist.ac.kr)

Mine for a Similar-Task Social Network
M. Song (m.song@unist.ac.kr)

Mine for a Subcontracting Social Network
M. Song (m.song@unist.ac.kr)

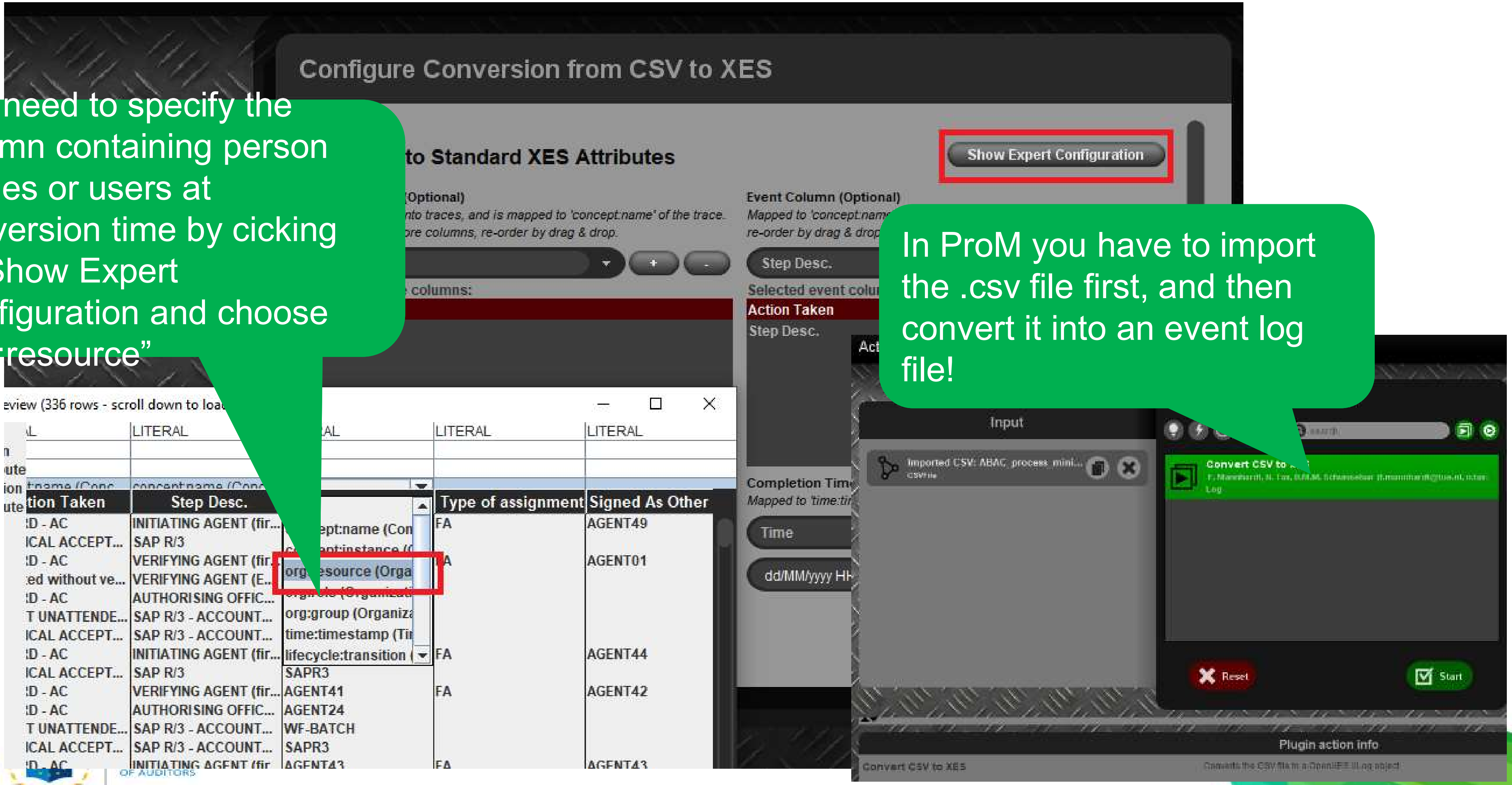
ing-Together Social Network

Social Network (WT)
SocialNetwork

Author: M. Song
Categories: Analytics

ProM is a framework with a large number of plugins. A social network graph is just another representation of the same data. In a conformance checking context, it is very useful to examine links within „populations of interest” that you isolate from the full population

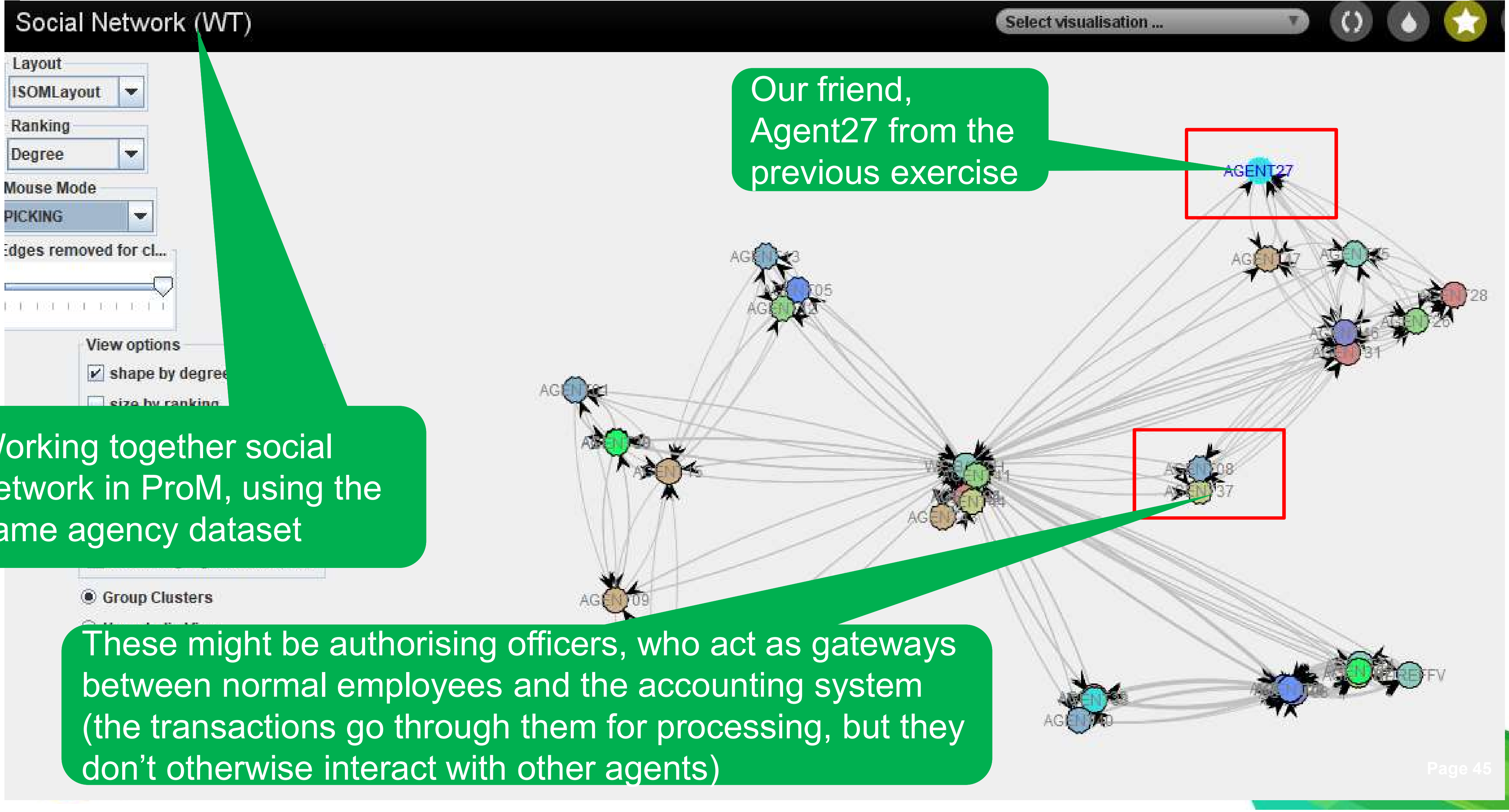
Import/conversion settings are important!!!



You need to specify the column containing person names or users at conversion time by cicking on Show Expert Configuration and choose “org:resource”

In ProM you have to import the .csv file first, and then convert it into an event log file!

Creating a social network based on the event log



Social Network (WT)

Layout: ISOMLayout

Ranking: Degree

Mouse Mode: PICKING

Edges removed for cl...

View options

☑ shape by degree

☐ size by ranking

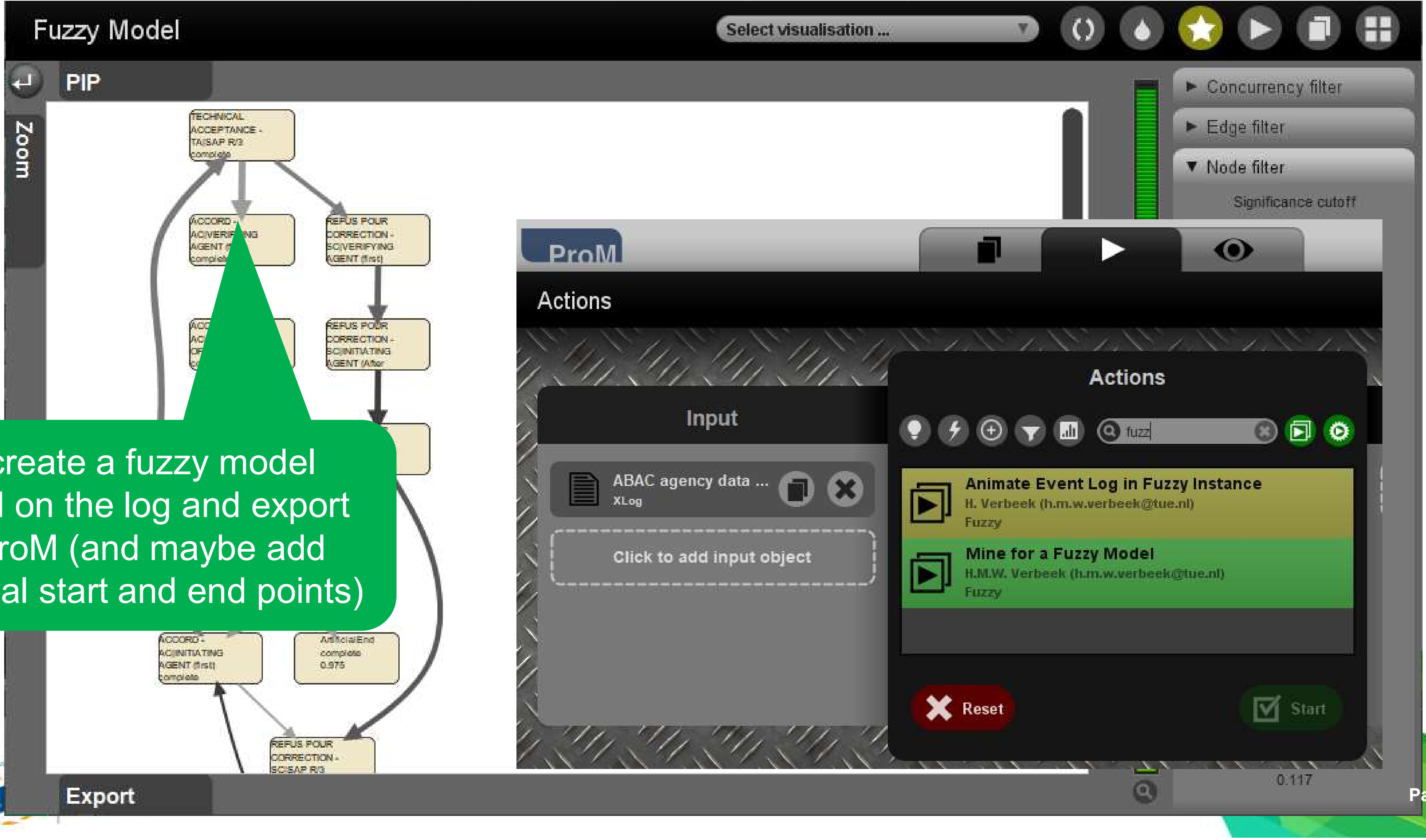
● Group Clusters

Our friend, Agent27 from the previous exercise

Working together social network in ProM, using the same agency dataset

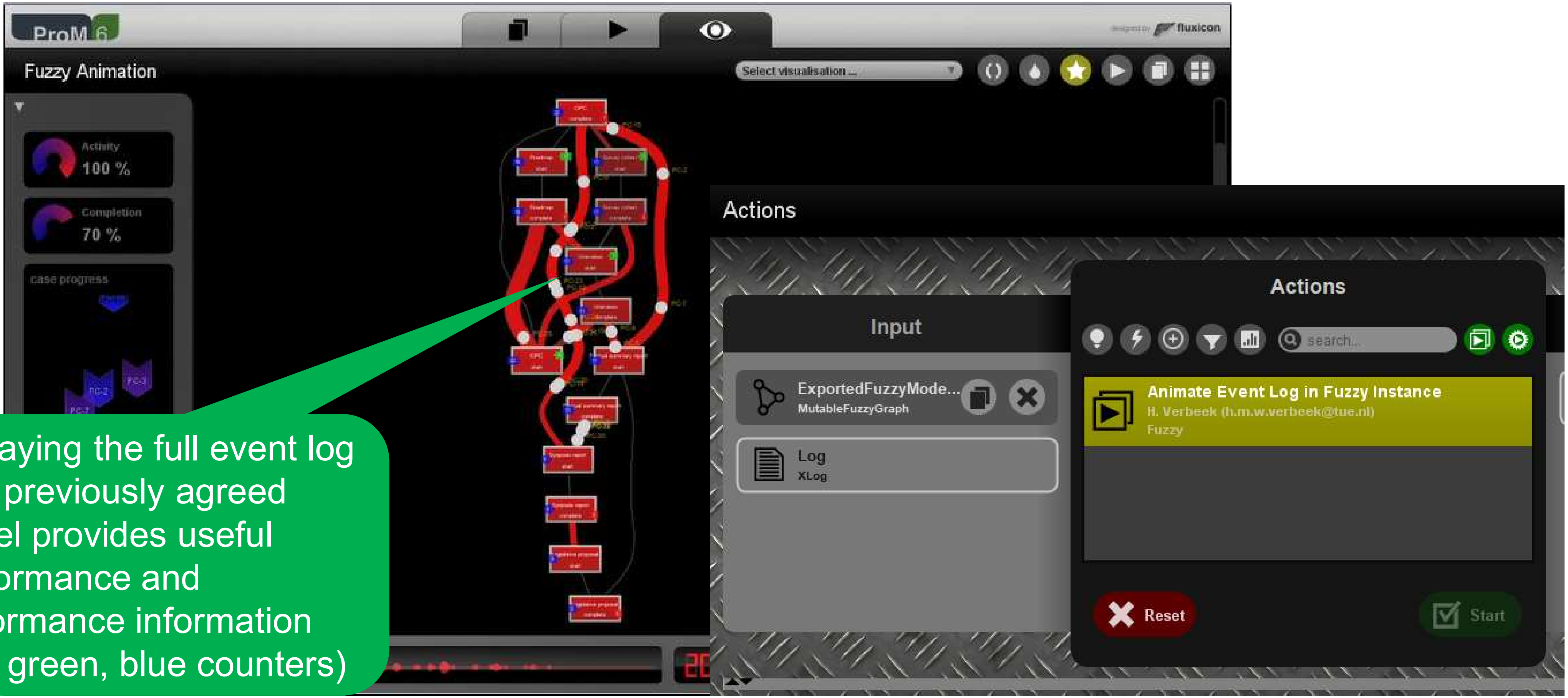
These might be authorising officers, who act as gateways between normal employees and the accounting system (the transactions go through them for processing, but they don't otherwise interact with other agents)

Replaying full event log on simplified process model


Fuzzy Model
Select visualisation ...
PIP
Zoom
TECHNICAL ACCEPTANCE - TA|SAP R/3 complete
ACCORD - AC|VERIFYING AGENT
REFUS POUR CORRECTION - SC|VERIFYING AGENT (first)
REFUS POUR CORRECTION - SC|INITIATING AGENT (After
ACCORD - AC|INITIATING AGENT (first) complete
ArtificialEnd complete 0.975
REFUS POUR CORRECTION - SC|SAP R/3
Export
Concurrency filter
Edge filter
Node filter
Significance cutoff
0.117
ProM
Actions
Input
ABAC agency data ...
XLog
Click to add input object
Actions
fuzz
Animate Event Log in Fuzzy Instance
H. Verbeek (h.m.w.verbeek@tue.nl)
Fuzzy
Mine for a Fuzzy Model
H.M.W. Verbeek (h.m.w.verbeek@tue.nl)
Fuzzy
Reset
Start

First create a fuzzy model based on the log and export it in ProM (and maybe add artificial start and end points)

Replaying full event log on simplified process model



Replaying the full event log on a previously agreed model provides useful conformance and performance information (red, green, blue counters)

Conformance checking in ProM

- There is a separate conformance checking plug-in in ProM that I have not covered in this presentation due to lack of time
- It works a bit differently than what I've shown in the previous slides
- Once a process model is agreed between auditor and auditee, the conformance checking results could be considered as audit evidence
- The ProM framework also allows you to write your own plugin in Java, so if you need checks that the system cannot currently do for you, you can implement them yourself in a plugin

Lessons learned

- In the worst case scenario, we can still extract useful process data from pdf files for prototyping (still needs manual touch-up, not suitable for production)
- We don't need much data to build the process model if the process is well-structured, 20 transactions gave us a useable model
- ProM/Disco is suitable for conformance checking, with certain limitations
- Some pre-processing is required to add event features (e.g. same agent with different logins/functions)
- No interactive display of non-conformances in Disco, need to use filters, Celonis might be a better option for this purpose (under investigation)

Mining Service Desk log files for process improvement

Process Mining Safari Part 3: being run over by an elephant

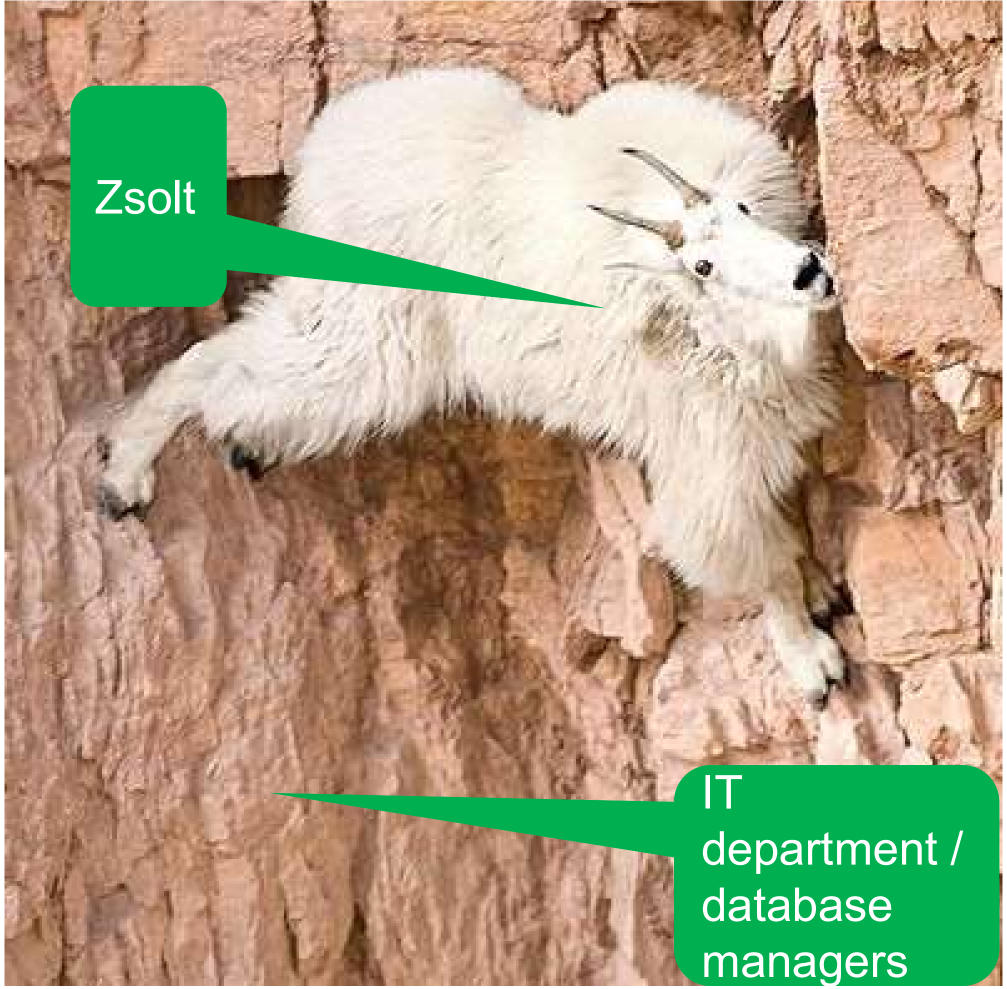
Business case

- Audit question: Explorative analysis of business processes with a view to potential process improvements
- Dataset: MS SQL server export from CA Service Desk Manager, containing only the keys, not the resolution for persons names, requestors etc.
- Business need: to gain experience with data extraction and process mining software, determine what can be done within PM software, what pre-processing steps may be required for client systems, etc.

My expectations

- Easy win, we just retired the system, accessing historical data on a non-live system should not be complicated
- It’s a service desk management system, processes should be well-defined
- I could present something useful relatively quickly to convince people about the advantages of process mining

Reality

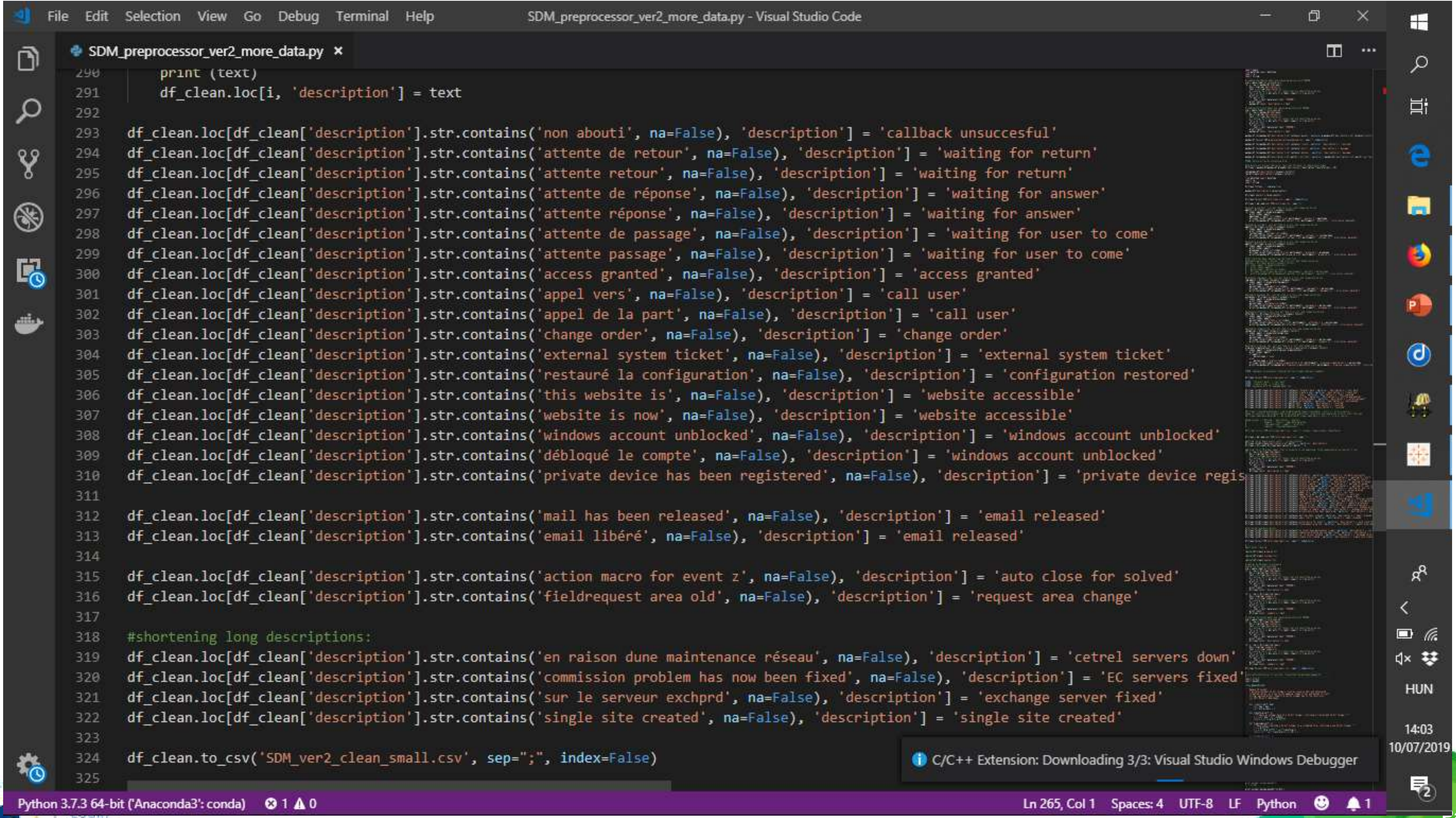

Zsolt
IT department / database managers

Data protection officer

The reality

- According to the DPO everything had to be anonymised, I did not extract ANY personal data so that we wouldn’t have to do extra administration
- Even the case IDs had to be encrypted, so they could not be traced back to actual persons
- Almost as many variants in the original as many cases
- French AND English event descriptions which had to be merged
- Emails had to be deleted using regexp, proper names had to be removed using Spacy EN/FR language models (NER)
- Explorative analysis had to determine the best splits of the data

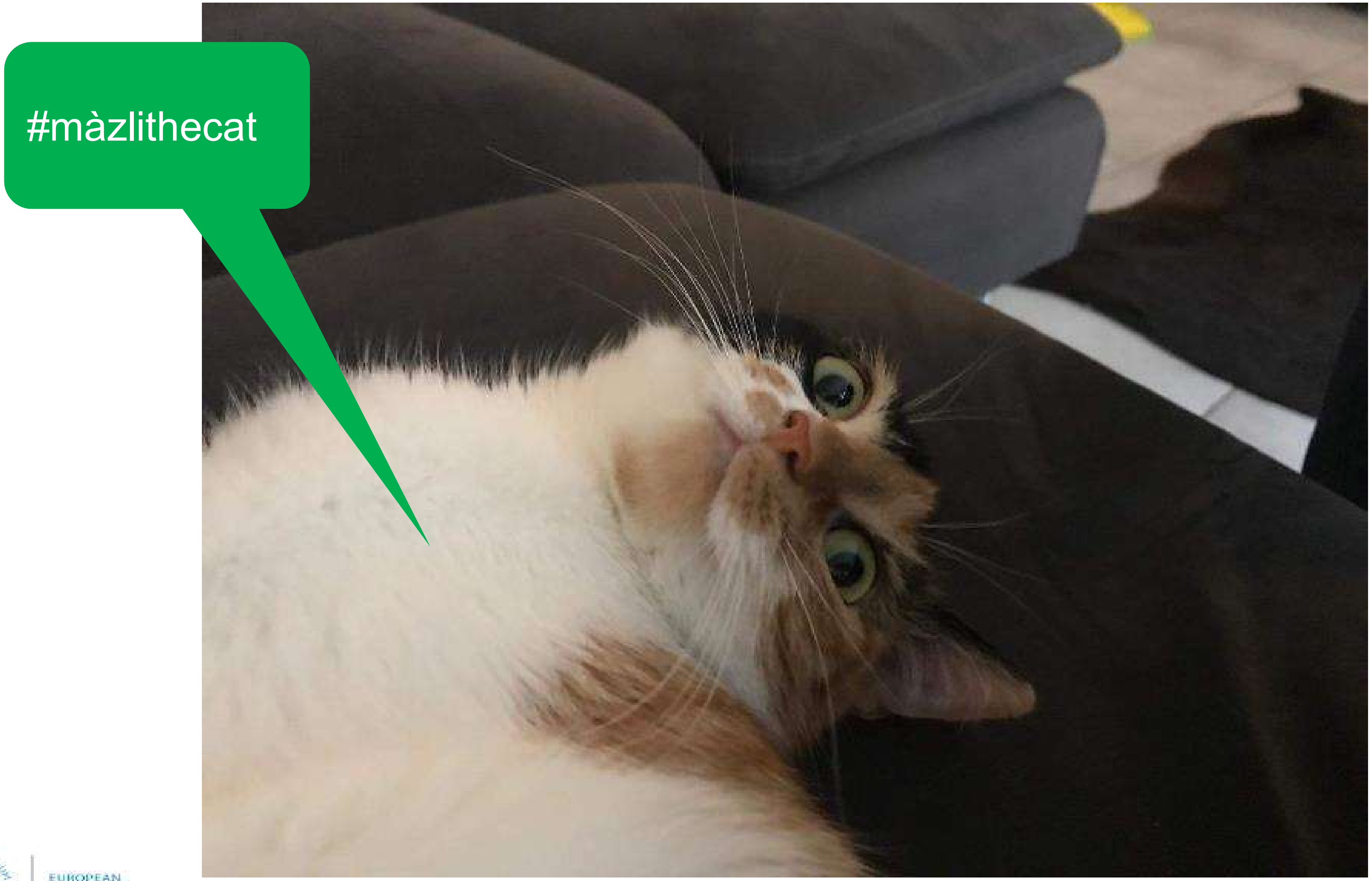
Pre-processing in Python, merging similar events


File
Edit
Selection
View
Go
Debug
Terminal
Help
SDM_preprocessor_ver2_more_data.py - Visual Studio Code
SDM_preprocessor_ver2_more_data.py
Python 3.7.3 64-bit ('Anaconda3': conda)
⊗ 1 ⚠ 0
Ln 265, Col 1
Spaces: 4
UTF-8
LF
Python
HUN
14:03
10/07/2019

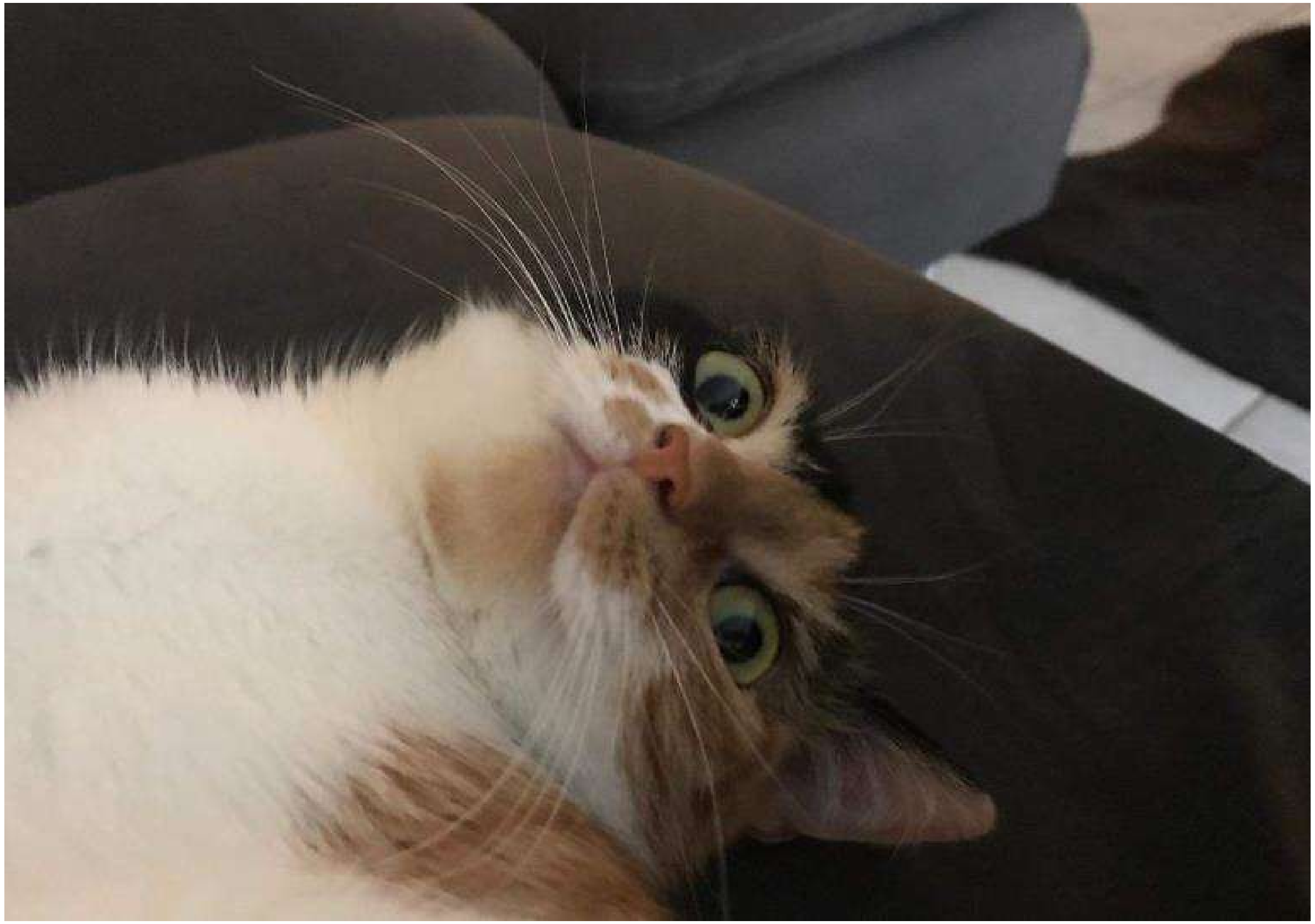
```python
    print (text)
    df_clean.loc[i, 'description'] = text

df_clean.loc[df_clean['description'].str.contains('non abouti', na=False), 'description'] = 'callback unsuccesful'
df_clean.loc[df_clean['description'].str.contains('attente de retour', na=False), 'description'] = 'waiting for return'
df_clean.loc[df_clean['description'].str.contains('attente retour', na=False), 'description'] = 'waiting for return'
df_clean.loc[df_clean['description'].str.contains('attente de réponse', na=False), 'description'] = 'waiting for answer'
df_clean.loc[df_clean['description'].str.contains('attente réponse', na=False), 'description'] = 'waiting for answer'
df_clean.loc[df_clean['description'].str.contains('attente de passage', na=False), 'description'] = 'waiting for user to come'
df_clean.loc[df_clean['description'].str.contains('attente passage', na=False), 'description'] = 'waiting for user to come'
df_clean.loc[df_clean['description'].str.contains('access granted', na=False), 'description'] = 'access granted'
df_clean.loc[df_clean['description'].str.contains('appel vers', na=False), 'description'] = 'call user'
df_clean.loc[df_clean['description'].str.contains('appel de la part', na=False), 'description'] = 'call user'
df_clean.loc[df_clean['description'].str.contains('change order', na=False), 'description'] = 'change order'
df_clean.loc[df_clean['description'].str.contains('external system ticket', na=False), 'description'] = 'external system ticket'
df_clean.loc[df_clean['description'].str.contains('restauré la configuration', na=False), 'description'] = 'configuration restored'
df_clean.loc[df_clean['description'].str.contains('this website is', na=False), 'description'] = 'website accessible'
df_clean.loc[df_clean['description'].str.contains('website is now', na=False), 'description'] = 'website accessible'
df_clean.loc[df_clean['description'].str.contains('windows account unblocked', na=False), 'description'] = 'windows account unblocked'
df_clean.loc[df_clean['description'].str.contains('débloqué le compte', na=False), 'description'] = 'windows account unblocked'
df_clean.loc[df_clean['description'].str.contains('private device has been registered', na=False), 'description'] = 'private device regis

df_clean.loc[df_clean['description'].str.contains('mail has been released', na=False), 'description'] = 'email released'
df_clean.loc[df_clean['description'].str.contains('email libéré', na=False), 'description'] = 'email released'

df_clean.loc[df_clean['description'].str.contains('action macro for event z', na=False), 'description'] = 'auto close for solved'
df_clean.loc[df_clean['description'].str.contains('fieldrequest area old', na=False), 'description'] = 'request area change'

#shortening long descriptions:
df_clean.loc[df_clean['description'].str.contains('en raison dune maintenance réseau', na=False), 'description'] = 'cetrel servers down'
df_clean.loc[df_clean['description'].str.contains('commission problem has now been fixed', na=False), 'description'] = 'EC servers fixed'
df_clean.loc[df_clean['description'].str.contains('sur le serveur exchprd', na=False), 'description'] = 'exchange server fixed'
df_clean.loc[df_clean['description'].str.contains('single site created', na=False), 'description'] = 'single site created'

df_clean.to_csv('SDM_ver2_clean_small.csv', sep=";", index=False)

```

C/C++ Extension: Downloading 3/3: Visual Studio Windows Debugger

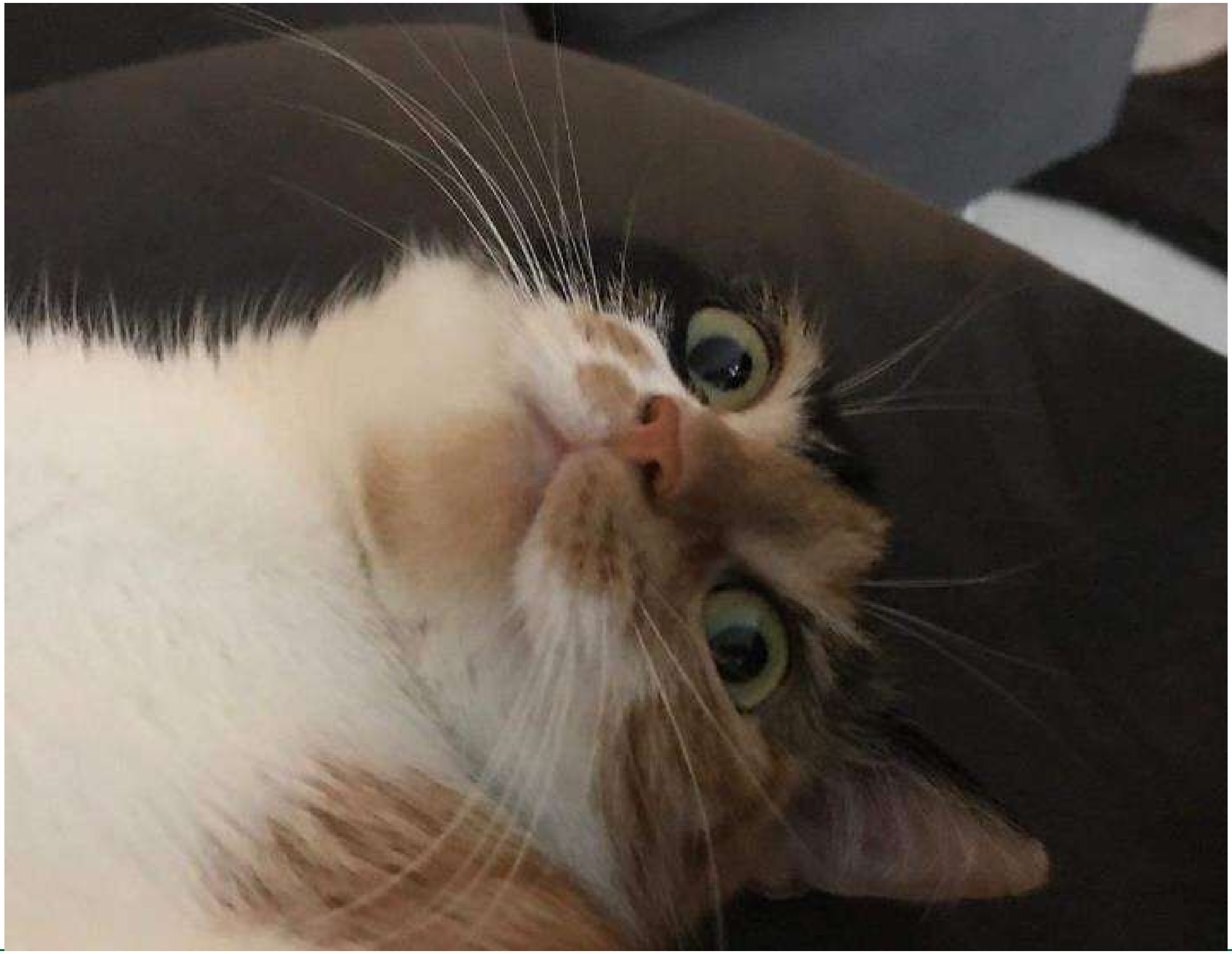
My face when I first saw the resulting process graph



My face when I first saw the resulting process graph



My face when I first saw the resulting process graph

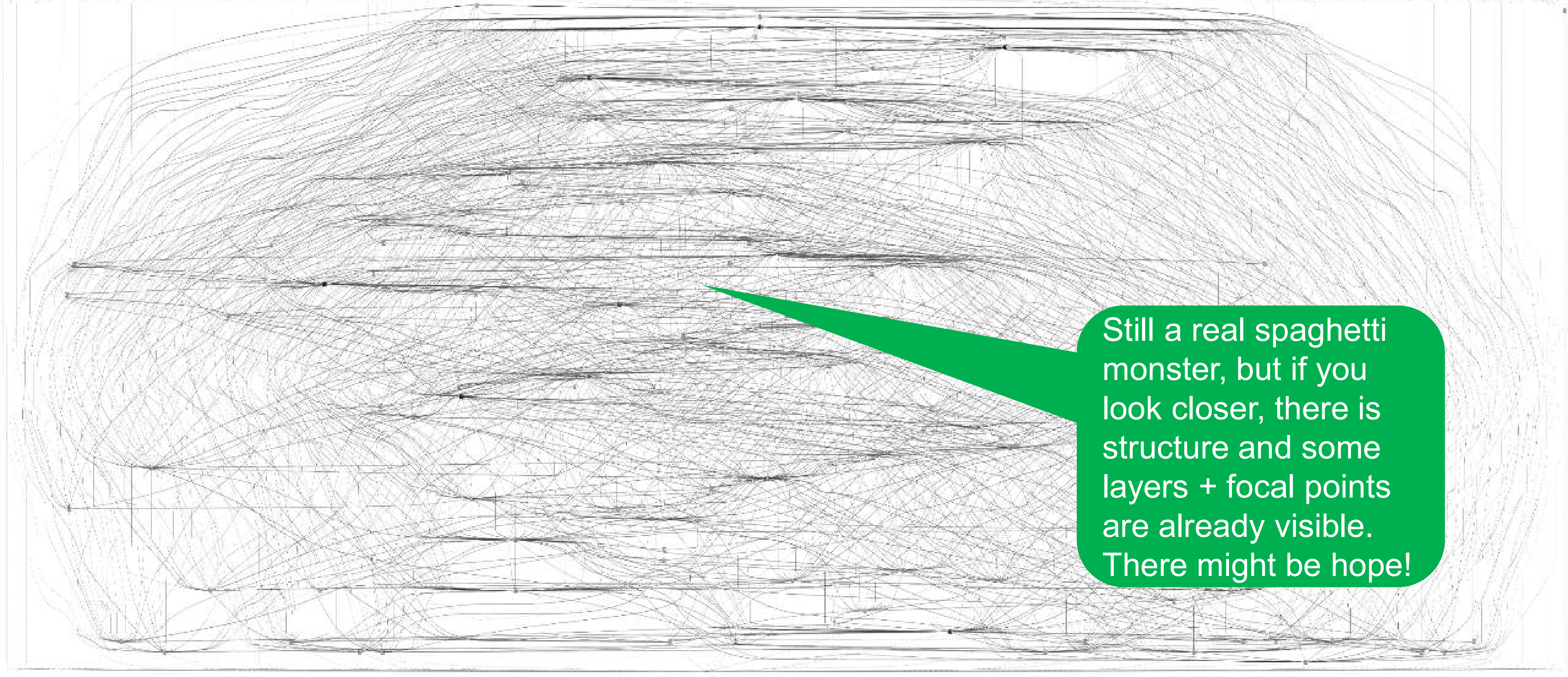


My face when I first saw the resulting process graph

My face when I first saw the resulting process graph

My face when I first saw the resulting process graph

The process graph derived from the pre-processed „simplified” log



Still a real spaghetti monster, but if you look closer, there is structure and some layers + focal points are already visible. There might be hope!

How to make lasagne out of spaghetti

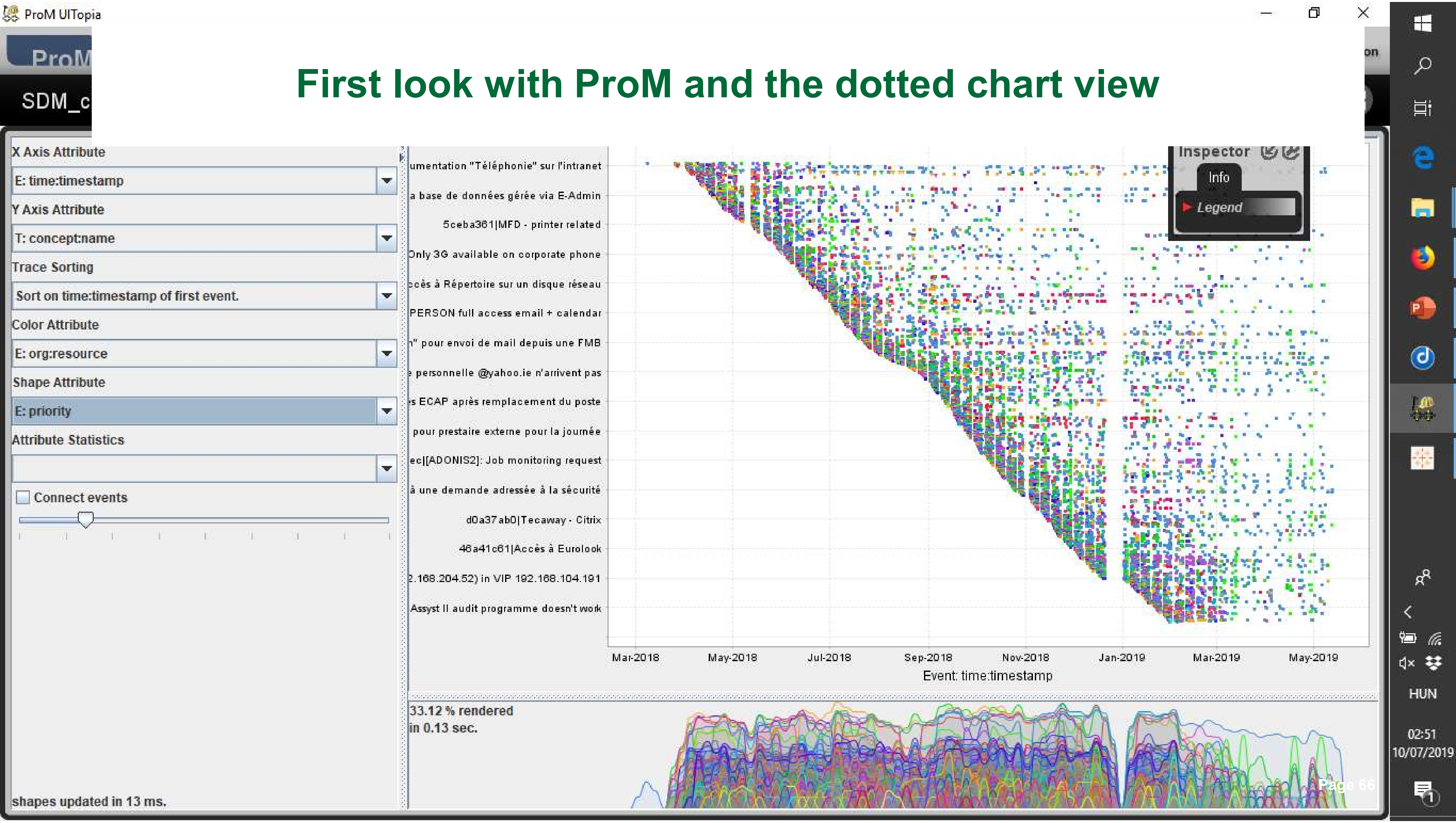
- It might as well be as difficult as turning water into wine
- Slice and dice! Explore the dotted chart, correlations, histograms, patterns.
- It might be the case that the log contains not one, but several processes, so you need to separate them!
- Talk to the business area and develop an intuition for recognising recurring patterns.

Good old Pareto principle

80/20 rule

20% of the cases/events cause 80% of the variance

First look with ProM and the dotted chart view



ProM UITopia

X Axis Attribute: E: time:timestamp

Y Axis Attribute: T: concept:name

Trace Sorting: Sort on time:timestamp of first event.

Color Attribute: E: org:resource

Shape Attribute: E: priority

Attribute Statistics:

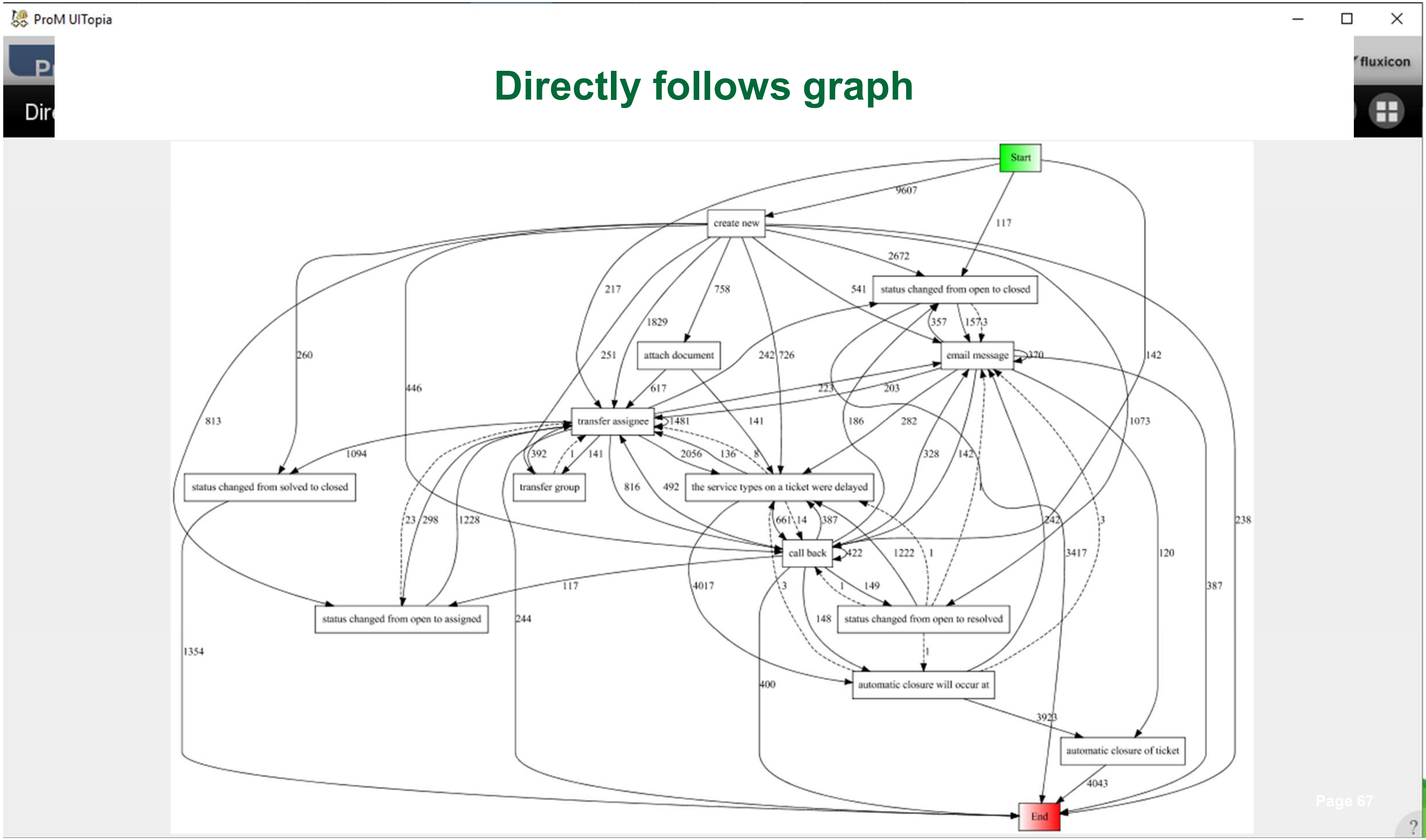
☐ Connect events

shapes updated in 13 ms.

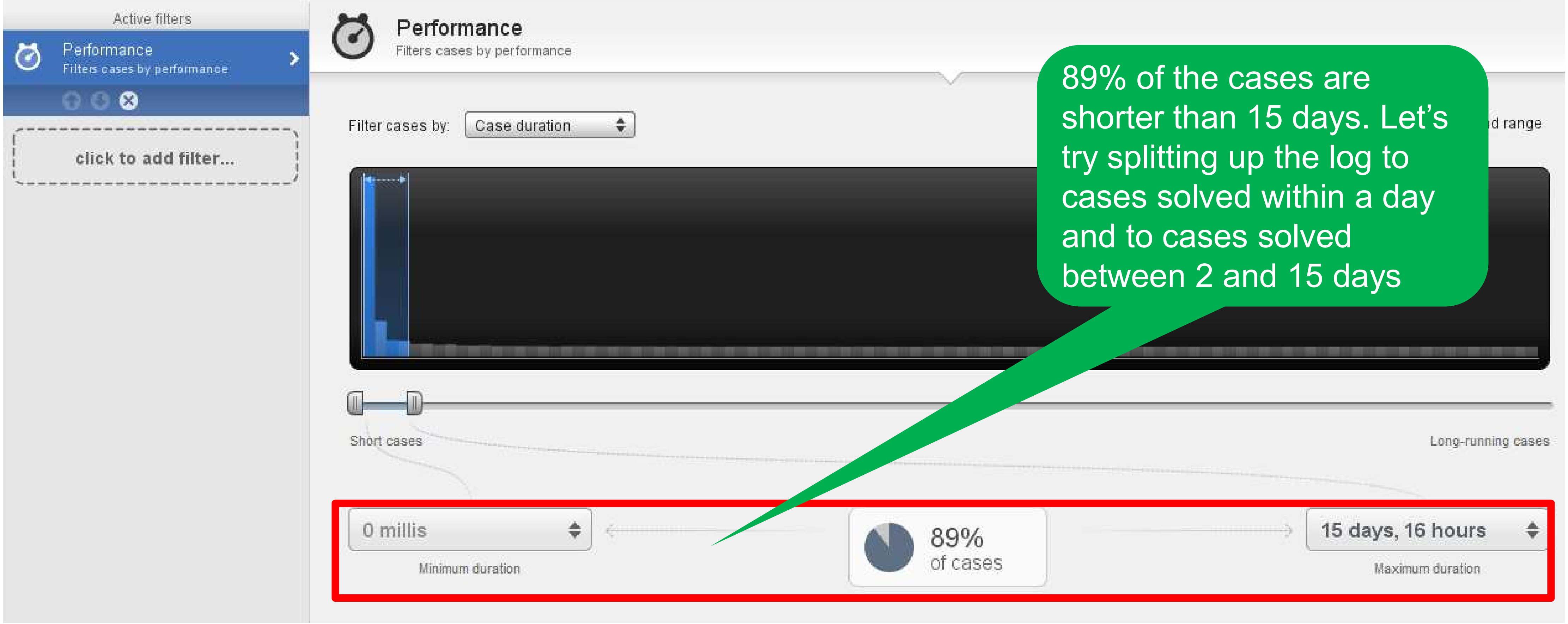
33.12 % rendered
in 0.13 sec.

ProM
SDM_c
Inspector
Info
Legend
umentation "Téléphonie" sur l'intranet
a base de données gérée via E-Admin
5ceba361|MFD - printer related
Only 3G available on corporate phone
ccès à Répertoire sur un disque réseau
PERSON full access email + calendar
n" pour envoi de mail depuis une FMB
e personnelle @yahoo.ie n'arrivent pas
s ECAP après remplacement du poste
pour prestaire externe pour la journée
ec|[ADONIS2]: Job monitoring request
à une demande adressée à la sécurité
d0a37ab0|Tecaway - Citrix
46a41c61|Accès à Eurolook
2.168.204.52) in VIP 192.168.104.191
Assyst II audit programme doesn't work
Mar-2018
May-2018
Jul-2018
Sep-2018
Nov-2018
Jan-2019
Mar-2019
May-2019
Event: time:timestamp
HUN
02:51
10/07/2019

Directly follows graph



Disco has easy to use, histogram based filters



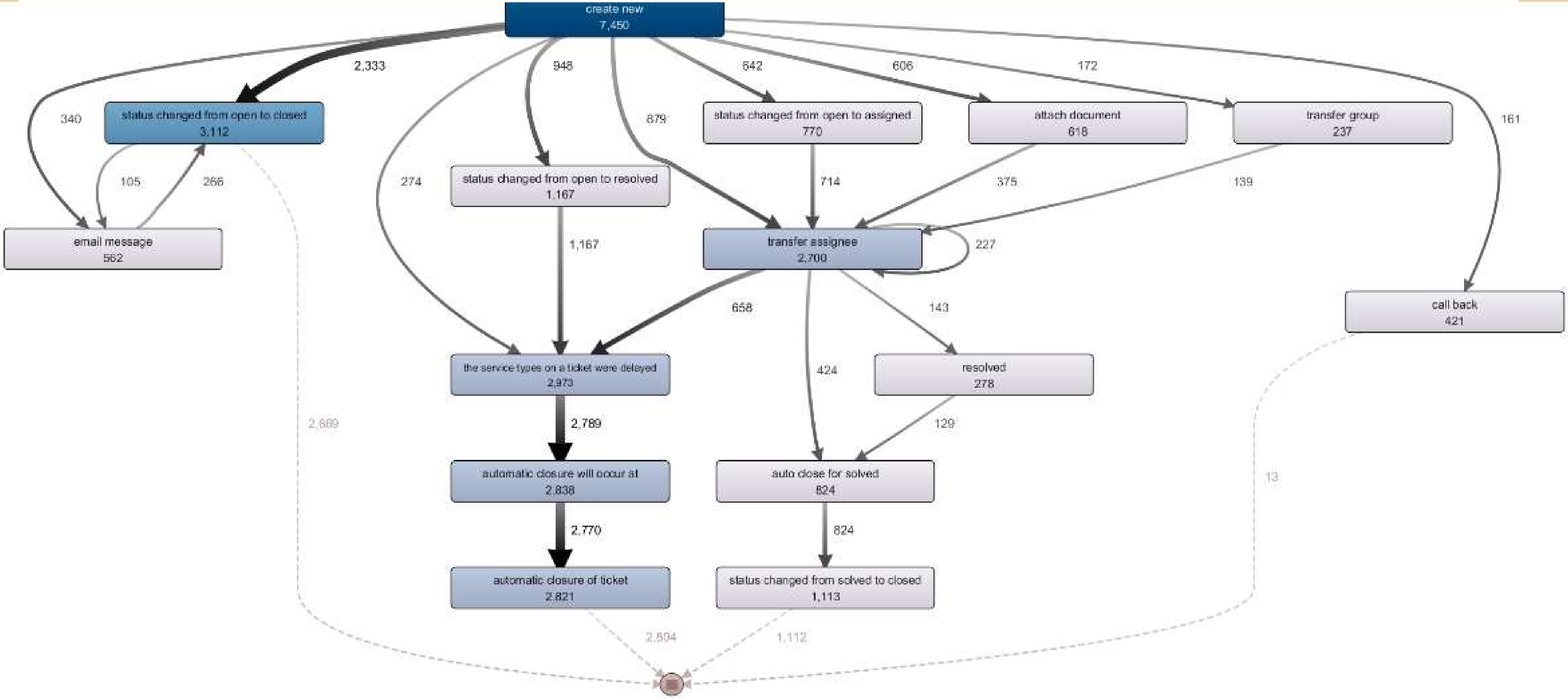
0 millis
Minimum duration

89%
of cases

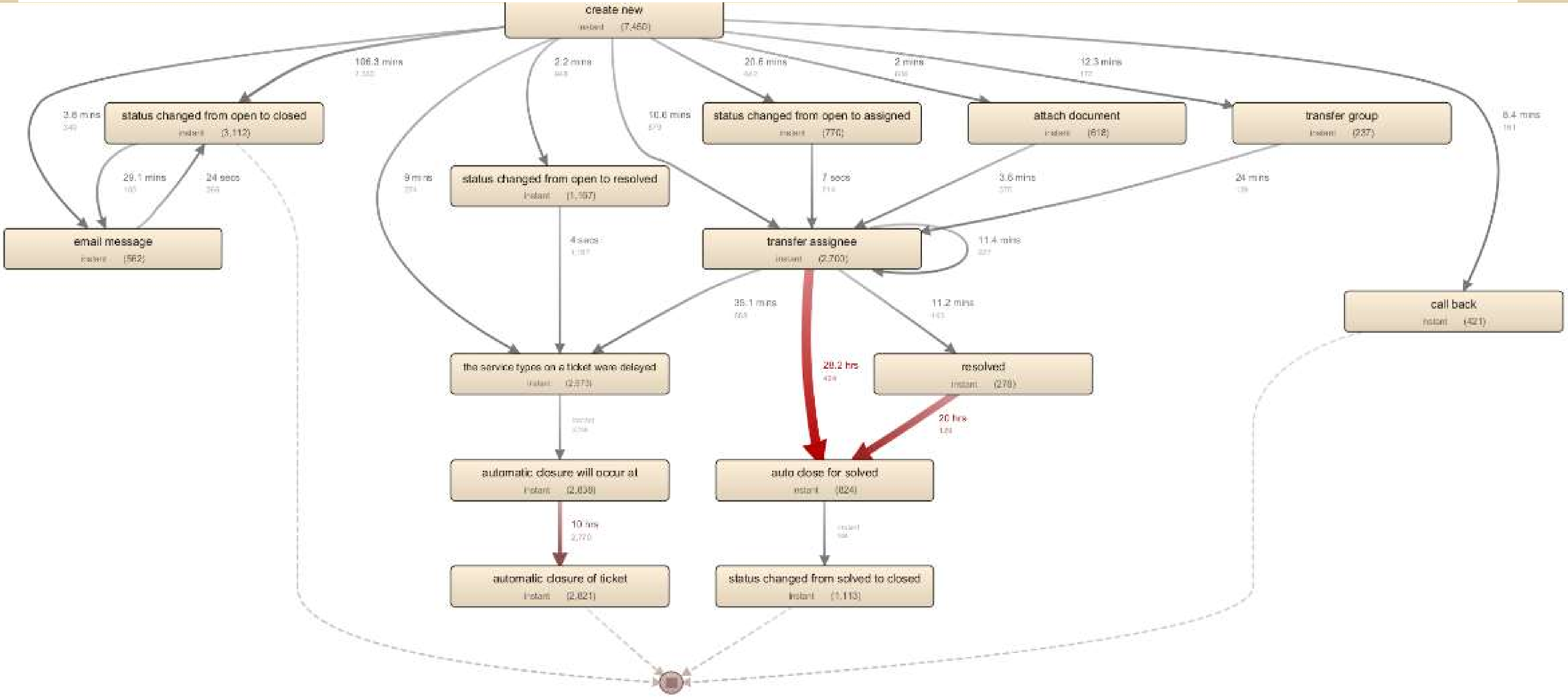
15 days, 16 hours
Maximum duration

89% of the cases are shorter than 15 days. Let’s try splitting up the log to cases solved within a day and to cases solved between 2 and 15 days

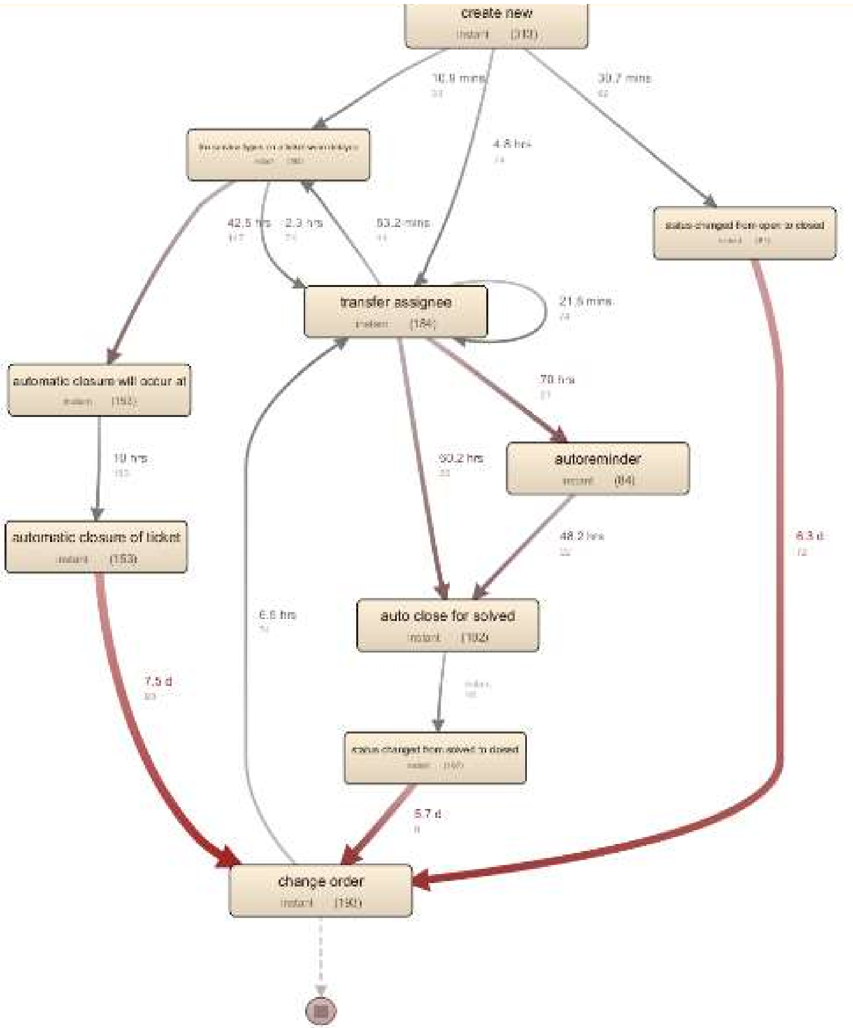
Split by duration (within a day)



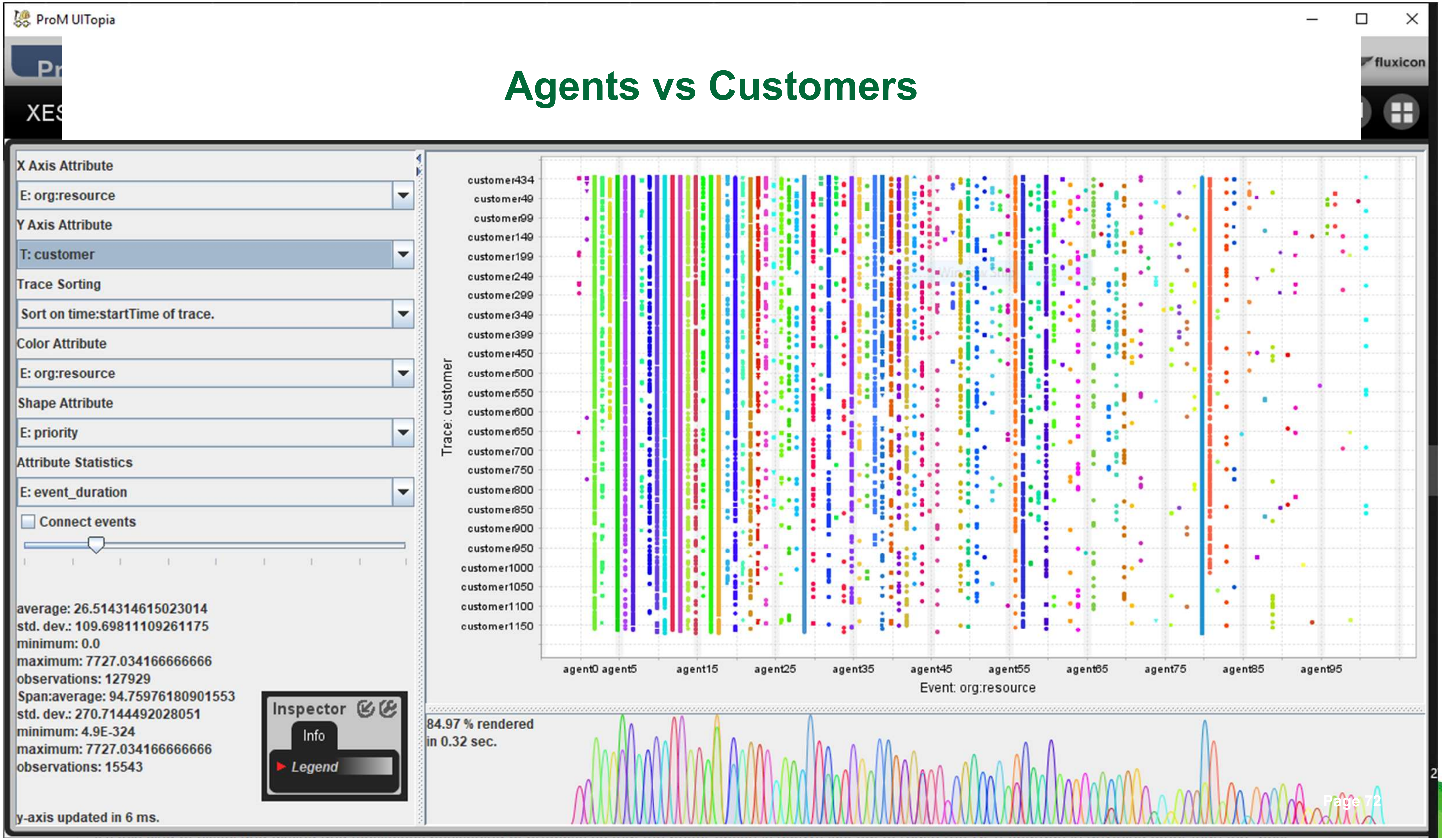
Split by duration (within a day)



Split by duration (2-15 days)



Agents vs Customers



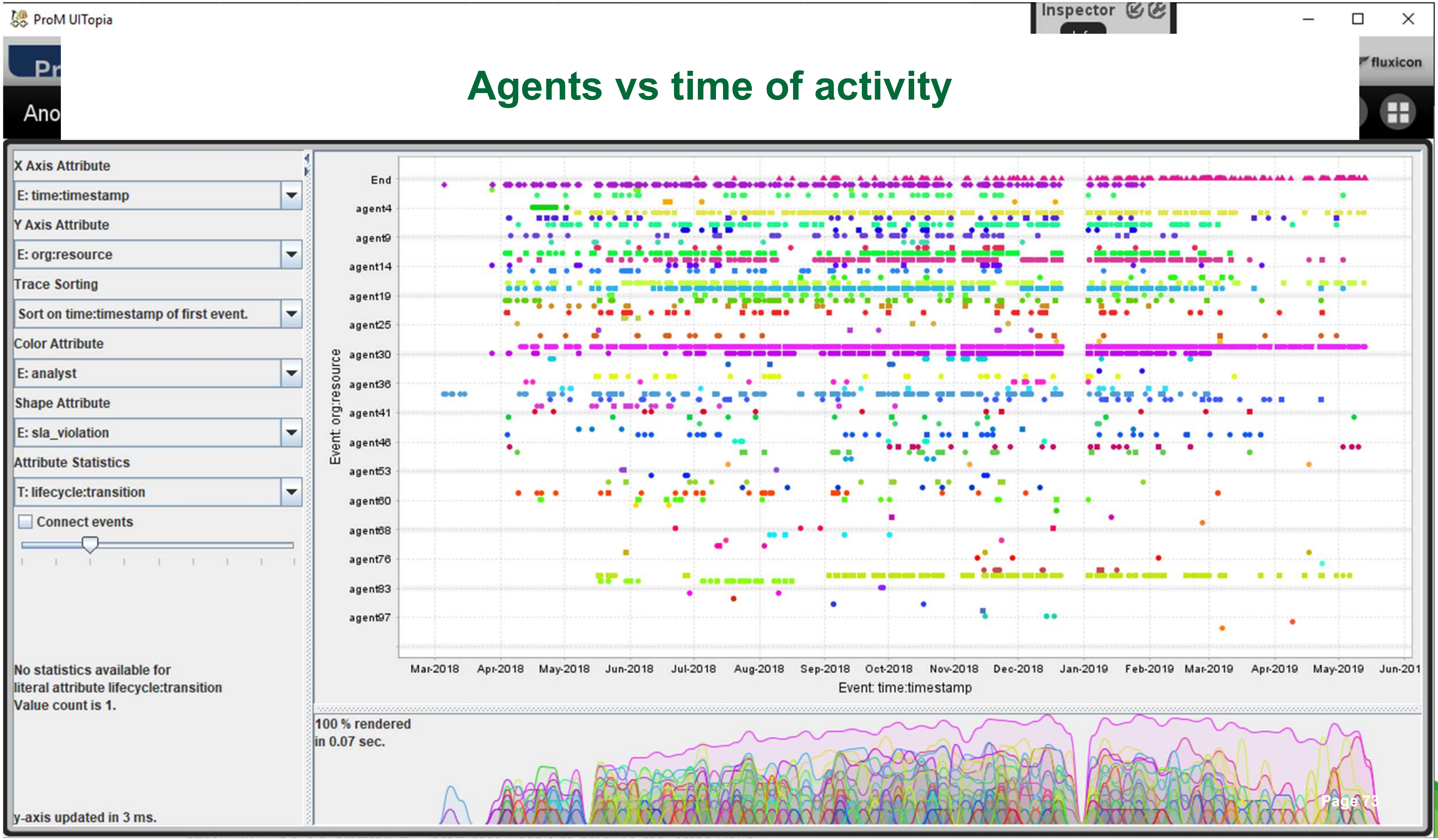
X Axis Attribute
E: org:resource
Y Axis Attribute
T: customer
Trace Sorting
Sort on time:startTime of trace.
Color Attribute
E: org:resource
Shape Attribute
E: priority
Attribute Statistics
E: event_duration
☐ Connect events

average: 26.514314615023014
std. dev.: 109.69811109261175
minimum: 0.0
maximum: 7727.034166666666
observations: 127929
Span:average: 94.75976180901553
std. dev.: 270.7144492028051
minimum: 4.9E-324
maximum: 7727.034166666666
observations: 15543

y-axis updated in 6 ms.

84.97 % rendered
in 0.32 sec.

Agents vs time of activity



X Axis Attribute

E: time:timestamp

Y Axis Attribute

E: org:resource

Trace Sorting

Sort on time:timestamp of first event.

Color Attribute

E: analyst

Shape Attribute

E: sla_violation

Attribute Statistics

T: lifecycle:transition

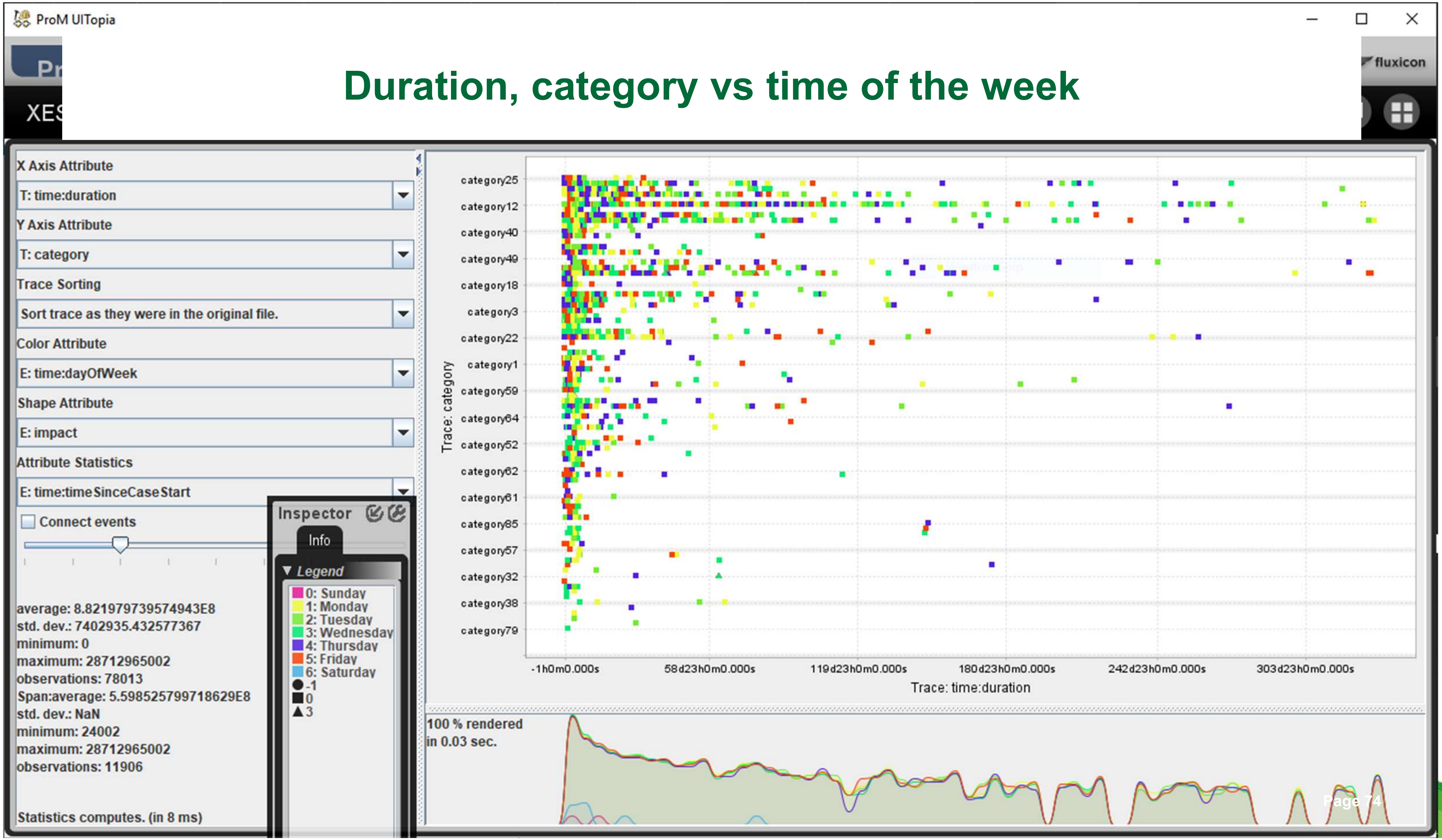
☐ Connect events

No statistics available for
literal attribute lifecycle:transition
Value count is 1.

y-axis updated in 3 ms.

100 % rendered
in 0.07 sec.

Duration, category vs time of the week



X Axis Attribute
T: time:duration

Y Axis Attribute
T: category

Trace Sorting
Sort trace as they were in the original file.

Color Attribute
E: time:dayOfWeek

Shape Attribute
E: impact

Attribute Statistics
E: time:timeSinceCaseStart

☐ Connect events

average: 8.821979739574943E8
std. dev.: 7402935.432577367
minimum: 0
maximum: 28712965002
observations: 78013
Span:average: 5.598525799718629E8
std. dev.: NaN
minimum: 24002
maximum: 28712965002
observations: 11906

Statistics computes. (in 8 ms)

Legend
- 🟪 0: Sunday
- 🟨 1: Monday
- 🟩 2: Tuesday
- 🟩 3: Wednesday
- 🟪 4: Thursday
- 🟧 5: Friday
- 🟦 6: Saturday
- ● -1
- ■ 0
- ▲ 3

100 % rendered
in 0.03 sec.

Lessons learned

- DO NOT assume that an event log will be easy to understand
- Use text mining tools to anonymise data, merge duplicate event names and quantify some of the unstructured textual data contained in description fields
- Be careful! Log content may contain personal data in fields where you are not expecting it! (e.g.: email text copied into description fields)
- Use your intuition and explore correlations and histograms
- The dotted chart view of ProM is very useful to get a „feel” for the data, and Disco has very nice histograms to help you slicing up the log data

Recommendations

- Start small: download a copy of ProM or Fluxicon Disco (demo) or register for Celonis cloud or contact Minit
- Start small: find an existing dataset that may be a good fit for process mining explorations
- Watch the Process Mining course on Coursera (https://www.coursera.org/learn/process-mining) and the ProM course on FutureLearn (https://www.futurelearn.com/courses/process-mining).
- Read Wil's book (https://www.springer.com/gp/book/9783662498507) and website (https://processmining.org)
- Read Fluxicon's online book (http://processminingbook.com/index.html)
- Secret tip: The Flux Capacitor blog (https://fluxicon.com/blog) especially the older entries between 2011-2014 contain a lot of tips and tricks

Recommendations

- Think big: if you are passionate you CAN make a difference
- Think big: this is a unique opportunity where public sector audit can be ahead of the private sector

- Act local: find out how automation/process mining can help your colleagues
- Act local: find out how your organisation may benefit from this

- Be realistic: it may or may not work out for you
- Be realistic: the technology might not be mature enough for your use case, but it might be overnight. Anticipate the change!

QUESTIONS?

Email: zsolt.varga@eca.europa.eu
LinkedIn: www.linkedin.com/in/zsvarga/
GitHub: github.com/zseebrz
Tableau Public: public.tableau.com/profile/zsolt.varga#!/