- Add an index page to navigate.
- Give credits everywhere.
- Format everything

BUSINESS PROCESS ANALYTICS

DataCamp course

Event data

- With the emergence of the Internet of Things, a lot of things around us are recording data about events that happen over time.
- As a result, the types of event data you can analyze is literally infinite.
- In this course you will learn about the different components of event data, and how to
- **Create**
- **Preprocess**
- **Analyze**
- the event data.

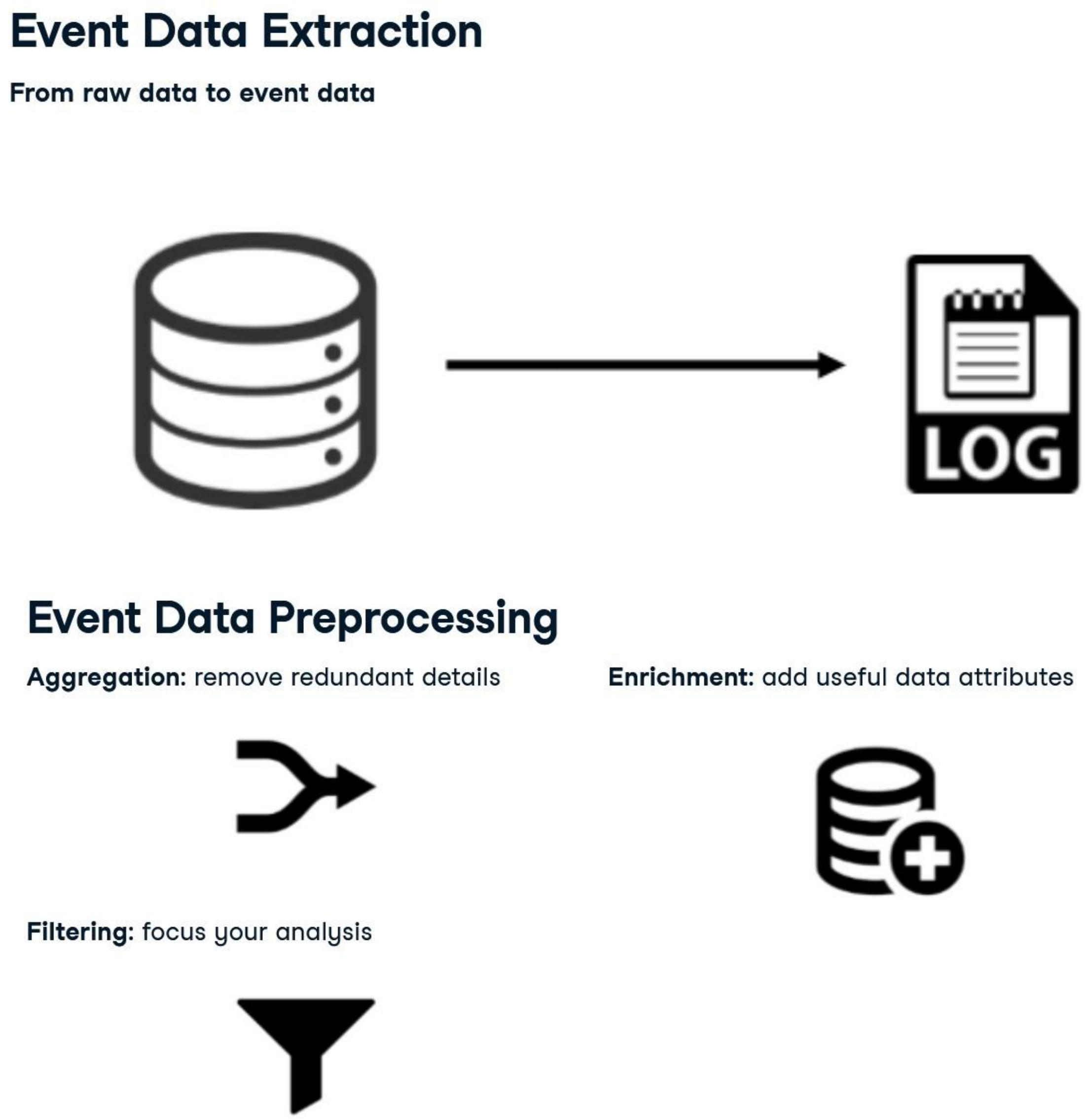
Event data

Event data

- Event data consists of three basic components:
- **the why,**
- **the what**
- **and the who.**
- an event is a recorded action of an activity (the what) occurring for an instance (the why) by a specific resource (the who).

Process analysis workflow

- Analyzing event data is an iterative process of three steps:
  – extraction,
  – processing
  – and analysis.



Event Data Extraction

From raw data to event data

Event Data Preprocessing

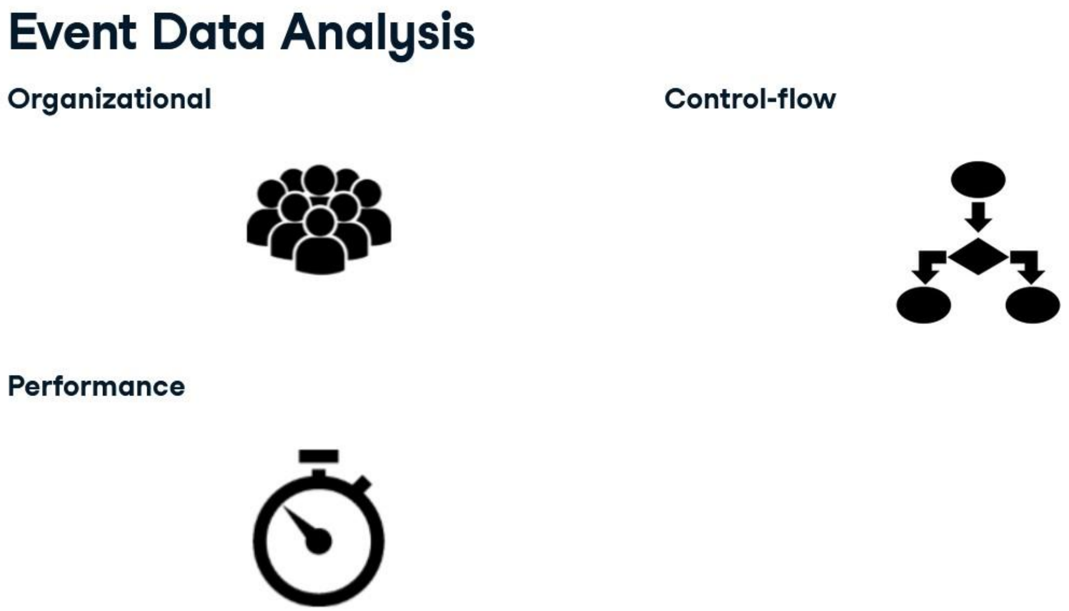
Aggregation: remove redundant details

Enrichment: add useful data attributes

Filtering: focus your analysis

Different perspectives after analysis

- Three perspectives can be distinguished.
- Firstly, the organizational perspective focuses on the actors.
- Secondly, the control-flow perspective, focusses on the flow and structuredness of the process.
- Finally, the performance perspective, focusses on time and efficiency.



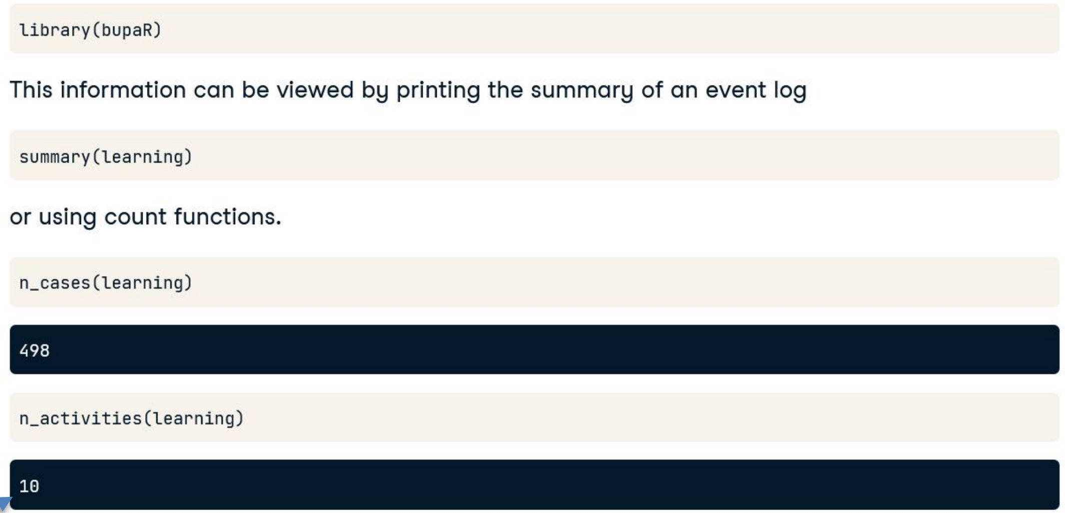
Event Data Analysis

Different perspectives after analysis

- Furthermore, we can also combine different perspectives, for example investigate whether there are links between actors and performance issues.
- Additional data attributes which are available, such as the cost of activities or types of customers, can also be included.

Exploring activities using bupaR

- **Example: Online learning**
- We are using a predefined package called bupaR which is used for process analysis.
- Basic process statistics can be inspected by printing the summary of an event log, or by using the count-functions such as displayed here.
- We can see that we have information on 498 students, who performed learning actions of 10 different types.



```
library(bupaR)
```

This information can be viewed by printing the summary of an event log

```
summary(learning)
```

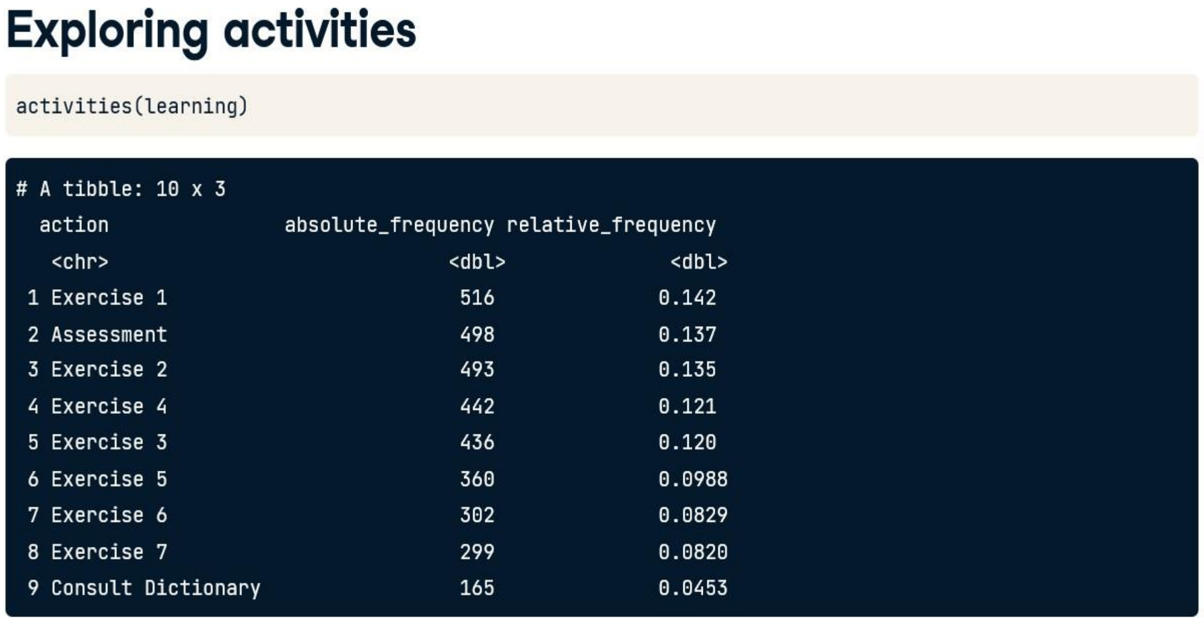
or using count functions.

```
n_cases(learning)
```

```
498
```

```
n_activities(learning)
```

```
10
```



Exploring activities

```
activities(learning)
```

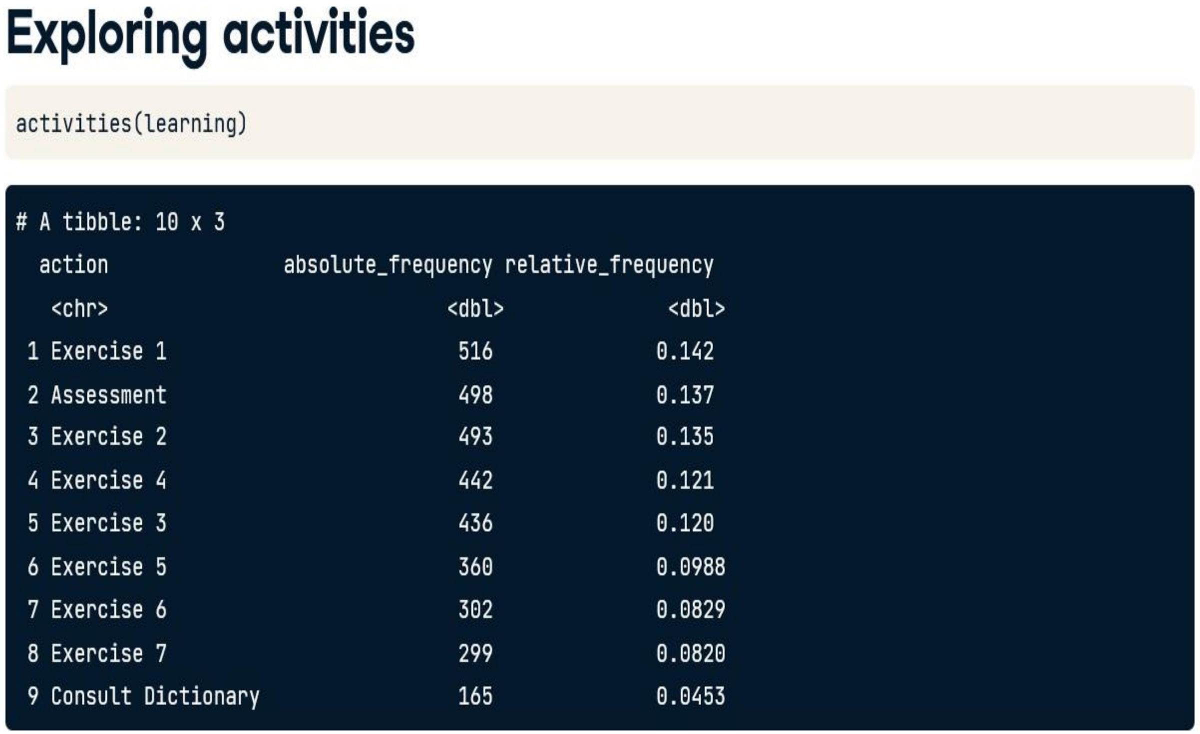
```
# A tibble: 10 x 3
  action               absolute_frequency relative_frequency
   <chr>                            <dbl>              <dbl>
 1 Exercise 1                         516             0.142
 2 Assessment                         498             0.137
 3 Exercise 2                         493             0.135
 4 Exercise 4                         442             0.121
 5 Exercise 3                         436             0.120
 6 Exercise 5                         360             0.0988
 7 Exercise 6                         302             0.0829
 8 Exercise 7                         299             0.0820
 9 Consult Dictionary                 165             0.0453
```

Exploring activities using bupaR

- **The activities are one of the most important characteristics** of a process, as they describe both the actions which are executed, and in which order this happens.
- More than anything else, activities define the process.
- Continuing the previous example, there are 10 different activities.
- We can retrieve the different types using the activity labels_function.
- We can see that there are 7 exercises and 1 assessment.
- Furthermore, students can also consult a dictionary and some theory pages.

Exploring activities using bupaR

- If we want more information on the activities, we can use the activities function.
- This will show us the number of times each of them occurs.
- In this example, exercise 1 has been done the most.
- In fact, the frequency is even higher than the number of students, which indicate that this exercise has been performed more than once by some students.
- On the other hand, consulting the dictionary or the theory pages has been done the least by the students.



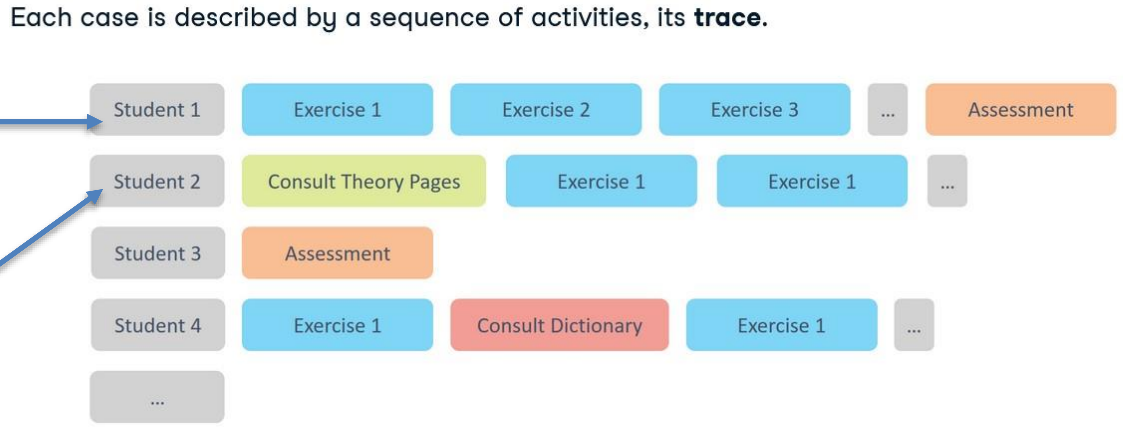
Exploring activities

```
activities(learning)
```

```
# A tibble: 10 x 3
  action             absolute_frequency relative_frequency
   <chr>                          <dbl>              <dbl>
 1 Exercise 1                       516             0.142
 2 Assessment                       498             0.137
 3 Exercise 2                       493             0.135
 4 Exercise 4                       442             0.121
 5 Exercise 3                       436             0.120
 6 Exercise 5                       360             0.0988
 7 Exercise 6                       302             0.0829
 8 Exercise 7                       299             0.0820
 9 Consult Dictionary               165             0.0453
```

Exploring the sequence of activities

- The sequence in which the activities occurs is called the trace of the case.
- A list of the traces can be retrieved with the `traces` function, or they can be visualized using the `trace_explorer`.
- For student 1, we see a very structured path, progressing through the different exercises, and finally doing the assessment.
- However, student 2 starts with looking at the theory pages, before doing exercise 1, which he executes two times in a row.



Each case is described by a sequence of activities, its **trace**.

Student 1	Exercise 1	Exercise 2	Exercise 3	...	Assessment
Student 2	Consult Theory Pages	Exercise 1	Exercise 1	...	
Student 3	Assessment				
Student 4	Exercise 1	Consult Dictionary	Exercise 1	...	
...					

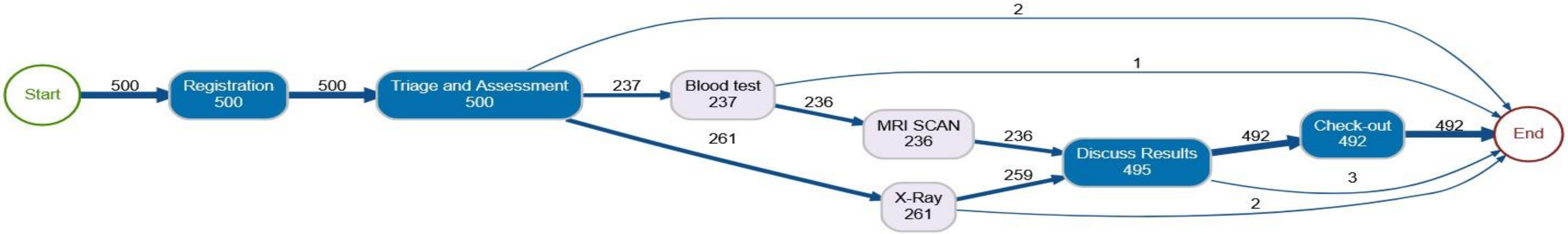
Process maps

Process maps

Another way to visualize processes is by constructing a process map.

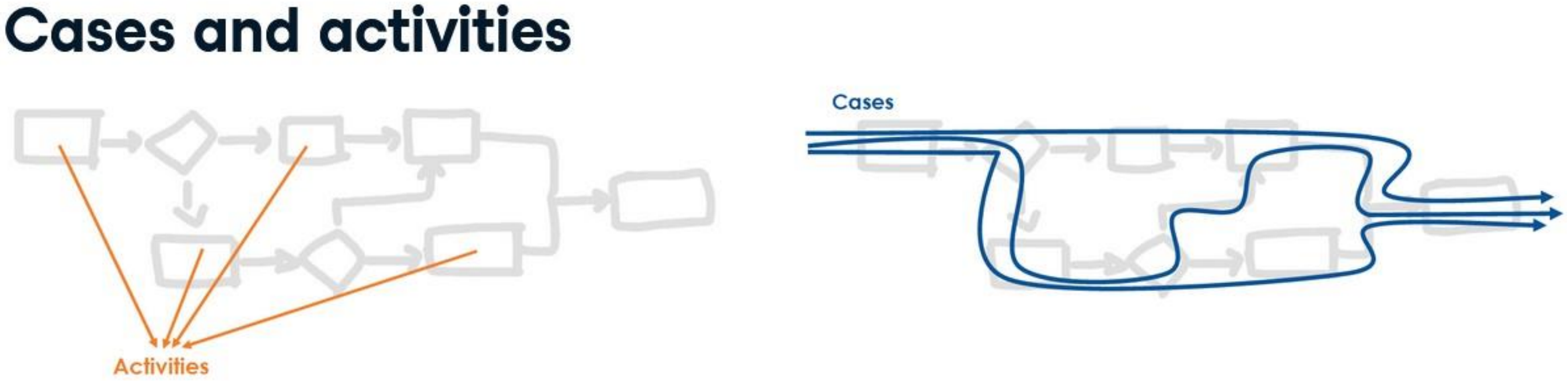
A process map is a directed graph that shows the activities of the process and the flows between them.

The colors of the nodes and the thickness of the arrows indicate the most frequent activities and process flows.



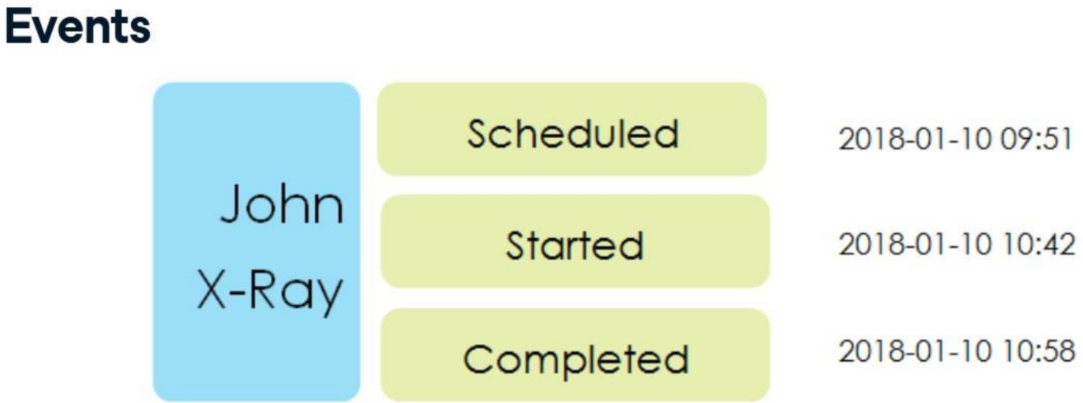
Components of process data

- So far we’ve covered two components of the process data.



Components of process data

- **Activity instances:**
- When a specific action takes place for a specific case, this is called an activity instance.
- In this example, we see three activity instances, containing two different activities: Registration and X-ray, and two different patients: Emily and John. In this example, we can see the time at which both the registration and the X-ray started.

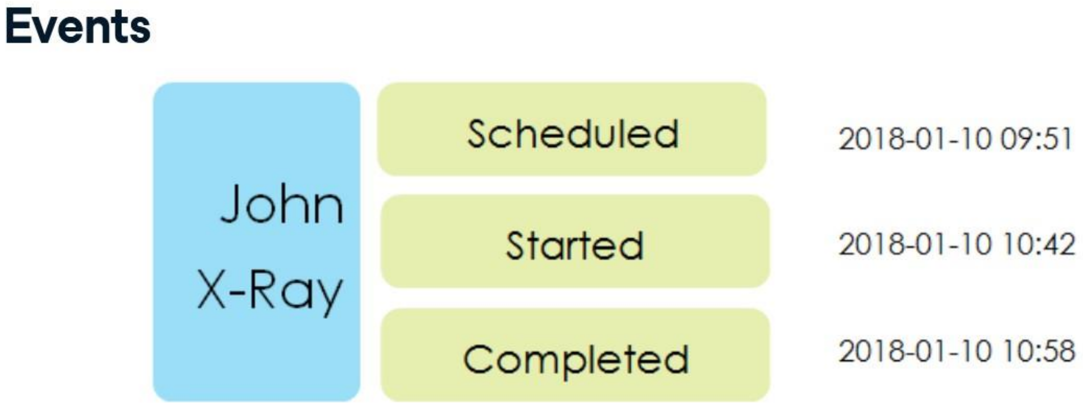


Events

John X-Ray	Scheduled	2018-01-10 09:51
	Started	2018-01-10 10:42
	Completed	2018-01-10 10:58

Components of process data

- **Events:**
- If we take a closer look at the X-Ray activity execution for John, we would see that it was scheduled during registration at 9:51.
- It actually started at 10:42, and was finished at 10:58. Each of these time recordings is called an event.

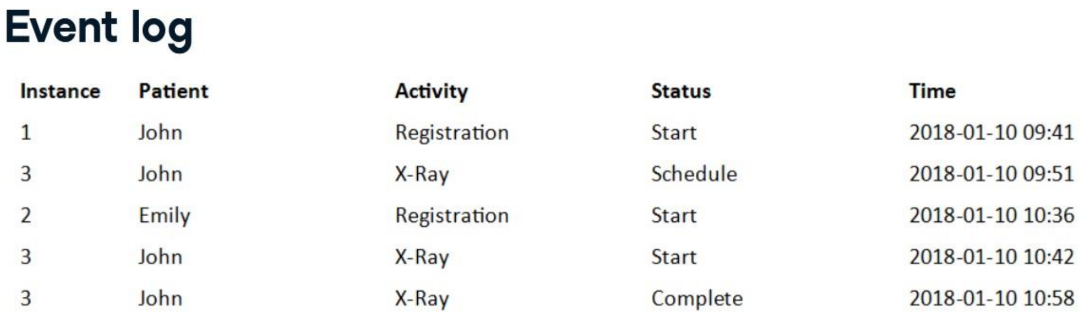


Events

John X-Ray	Scheduled	2018-01-10 09:51
	Started	2018-01-10 10:42
	Completed	2018-01-10 10:58

Components of process data

- **Event log**
- A log book of events is called an event log.
- For our current example, the event log looks as shown here.
- There are two registration instances, for which only a start event is recorded (the arrival of the patient), and there is 1 X-ray instance which consists of 3 events.

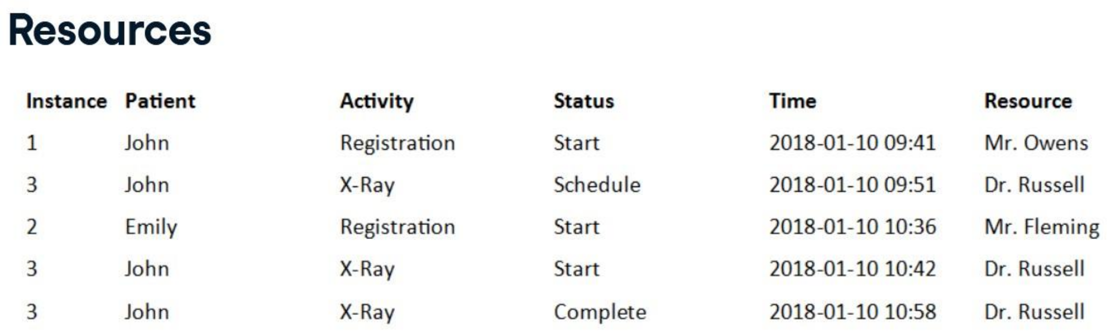


Event log

Instance	Patient	Activity	Status	Time
1	John	Registration	Start	2018-01-10 09:41
3	John	X-Ray	Schedule	2018-01-10 09:51
2	Emily	Registration	Start	2018-01-10 10:36
3	John	X-Ray	Start	2018-01-10 10:42
3	John	X-Ray	Complete	2018-01-10 10:58

Components of process data

- **Resources**
- Another component of process data are the resources.
- Resources are the actors in the process.
- In our example, there are three resources: Mr. Owens and Mr. Fleming who performed the registration of John and Emily, respectively, and Dr. Russell who performed the X-ray.



Resources

Instance	Patient	Activity	Status	Time	Resource
1	John	Registration	Start	2018-01-10 09:41	Mr. Owens
3	John	X-Ray	Schedule	2018-01-10 09:51	Dr. Russell
2	Emily	Registration	Start	2018-01-10 10:36	Mr. Fleming
3	John	X-Ray	Start	2018-01-10 10:42	Dr. Russell
3	John	X-Ray	Complete	2018-01-10 10:58	Dr. Russell

Different perspectives after analysis

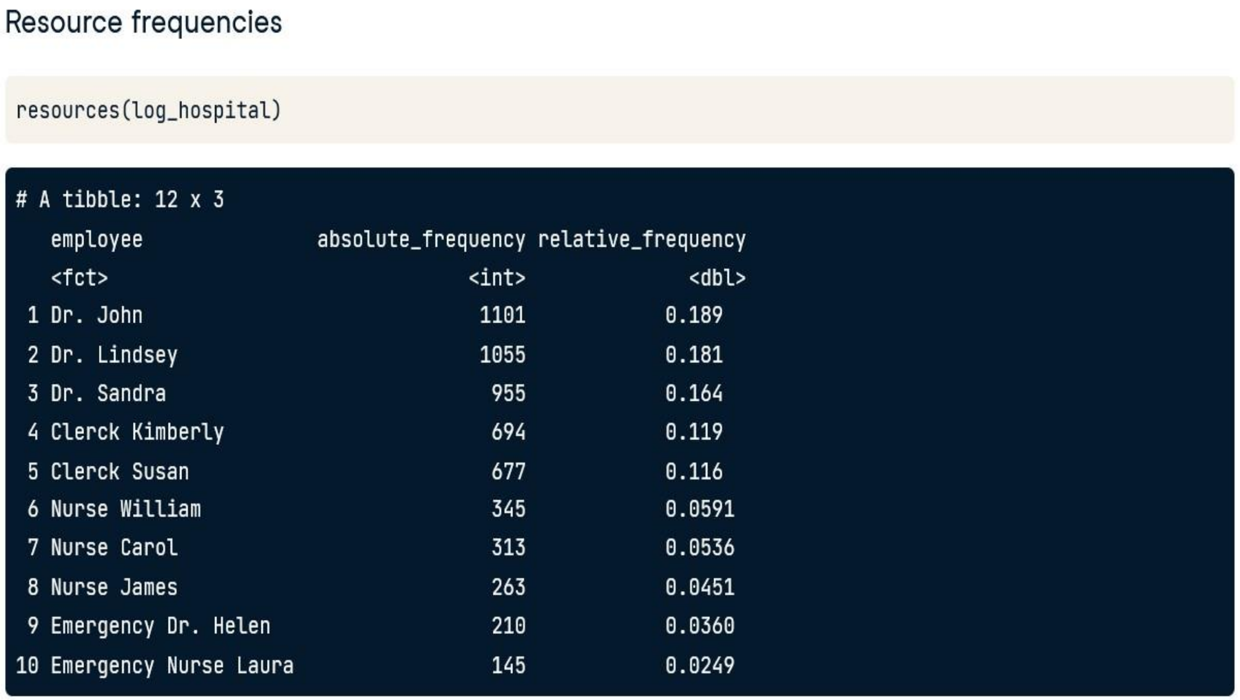
- Organizational analysis

Organizational analysis

Data: Hospital process

Organizational analysis

- **Who executes the work?**
- In order to know who executes the work, we can take a look at the resource labels, using the resource_labels function.
- We can see that there are 12 resources in this example process: doctors, nurses, clercks and an emergency crew.



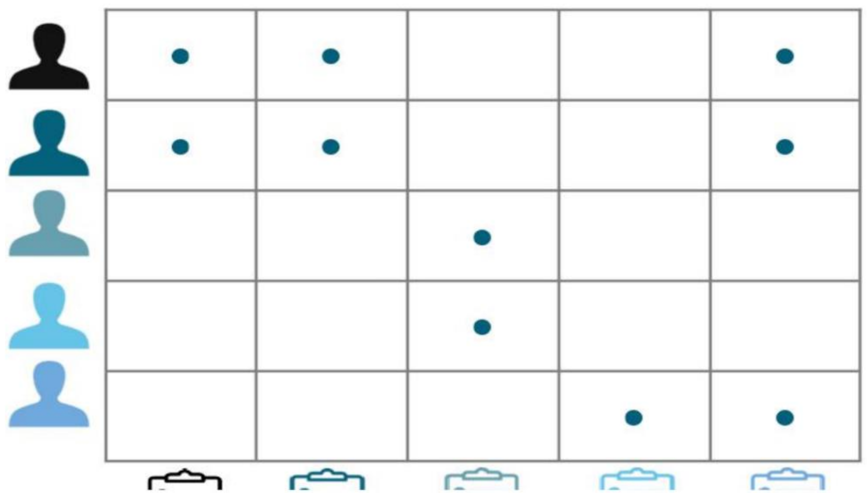
Resource frequencies

```
resources(log_hospital)
```

```
# A tibble: 12 x 3
   employee              absolute_frequency relative_frequency
   <fct>                              <int>              <dbl>
 1 Dr. John                            1101             0.189
 2 Dr. Lindsey                         1055             0.181
 3 Dr. Sandra                           955             0.164
 4 Clerck Kimberly                      694             0.119
 5 Clerck Susan                         677             0.116
 6 Nurse William                        345             0.0591
 7 Nurse Carol                          313             0.0536
 8 Nurse James                          263             0.0451
 9 Emergency Dr. Helen                  210             0.0360
10 Emergency Nurse Laura                145             0.0249
```

Organizational analysis

- **Resource activity matrix:**
- A resource-activity matrix shows what resource in your organization performs what task.

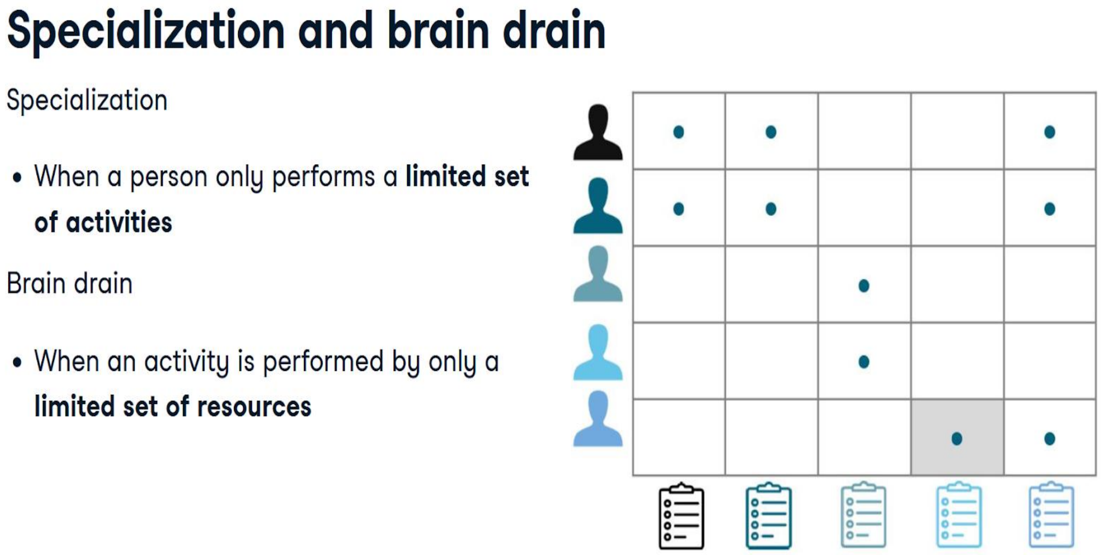


Resource-activity Matrix

●	●			●
●	●			●
		●		
		●		
			●	●

Organizational analysis

- **Specialization and Brain drain:**
- When a person performs relatively few activities, we can say that this person is specialized, like person 3 and 4 in this example.
- Knowing the specializations of your work-force is very important from a knowledge management perspective: knowing who to go to for a specific issue, or knowing who to enlist for specific trainings.



Specialization and brain drain

Specialization

- When a person only performs a **limited set of activities**

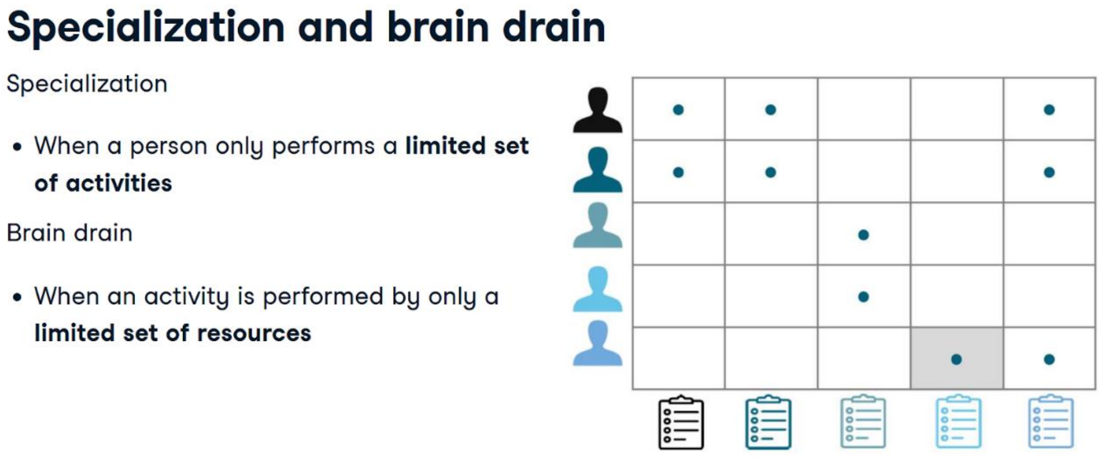
Brain drain

- When an activity is performed by only a **limited set of resources**

●	●			●
●	●			●
		●		
		●		
			●	●

Organizational analysis

- **Specialization and Brain drain:**
- On the other hand, it can also occur that only one person is in charge of a certain activity, like the forth activity in this example.
- This presents an important risk in terms of knowledge retention: if the person leaves the company, you might suffer an important loss of knowledge concerning this step of the process.



Specialization and brain drain

Specialization

- When a person only performs a **limited set of activities**

Brain drain

- When an activity is performed by only a **limited set of resources**

●	●			●
●	●			●
		●		
		●		
			●	●

Control flow analysis

- The structuredness of the process.
- When we look at structuredness, we often speak of "control-flow".
- Control-flow refers to the different successions of activities, like in the example we saw of the different students.
- Recall that each case can be expressed as a sequence of activities.
- Each unique sequence is called a trace or process variant.

Control flow analysis

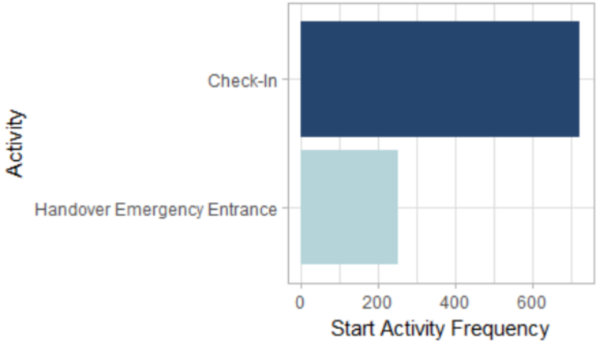
- There are several ways to look at the variants of the process.
- On one hand, we have several metrics which we can use to look at one specific aspect of the process:
  - start and end activities,
  - the distribution of the case length,
  - which activities are always present in a case and
  - which are exceptional

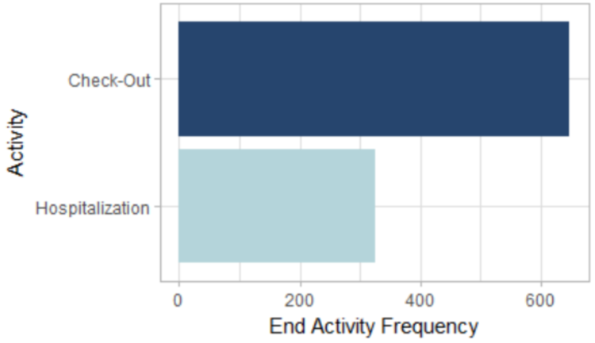
Visualizing the control-flows

- For each of the topics in the last slide, a metric similar to that of resource_frequency of can be calculated.
- There are various visual tools to look at control-flow patterns:
- Process map
- Trace explorer
- Precedence matrix.

Visualizing the control-flows

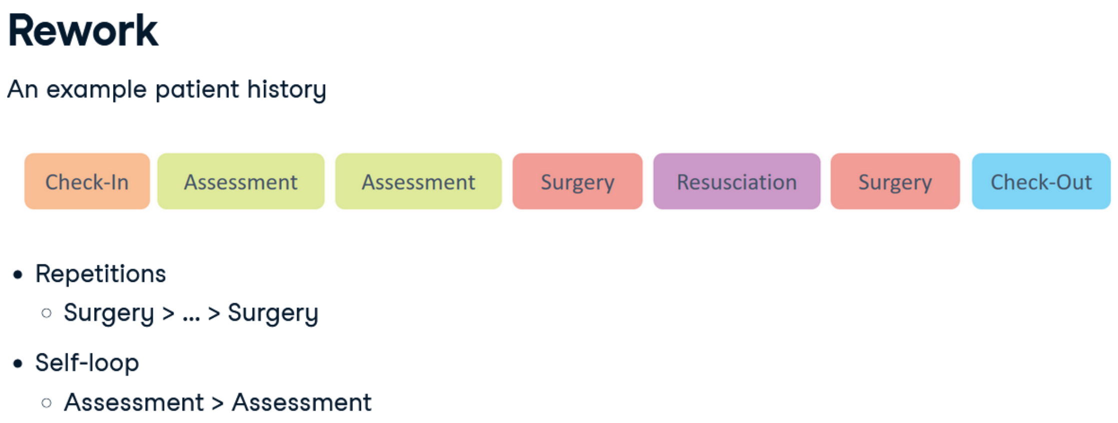
- We see that there are 2 entry points and 2 exit points in this process.
- Most patient journey's start with a check-in and end with a check-out.
- However, some are handed over by the emergency vehicle crew. Not all patients are check-out, as some of them need further treatment in the hospital.





Visualizing the control-flows

- Another aspect is **rework**.
- When we talk about rework, it means that some activities in the process are done multiple times for the same case, which often is a source of inefficiencies and waste.
- For example, consider this journey of a particular patient.
- We distinguish two types of rework:
- Firstly, repetitions, when an activity is repeated later on in the case, such as surgery in this example.
- Secondly, self-loops, when an activity is repeated immediately, such as assessment.



Rework

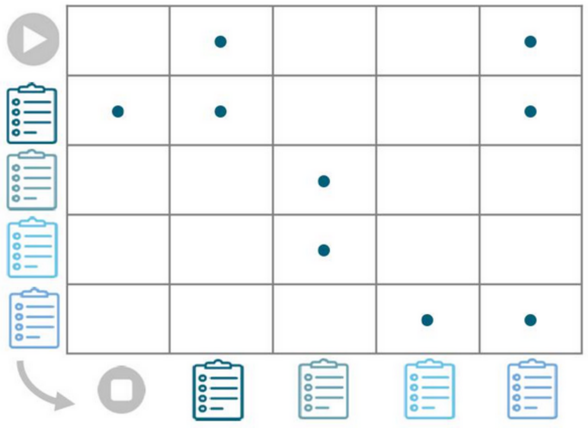
An example patient history

- Repetitions
  - Surgery > ... > Surgery
- Self-loop
  - Assessment > Assessment

Visualizing the control-flows

- As a visual, we can create precedence matrices of process logs.
- A precedence matrix shows the flows between activities in a more structured way compared to process map.
- It thus shows which activities precede other activities, which ones are at the start or the end, and which are the most important flows in the process.

Precedence matrix

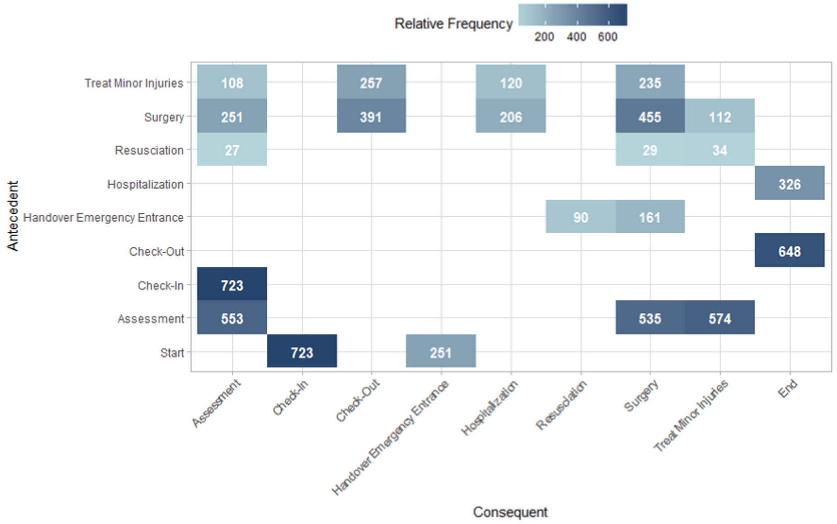


▶		●			●
📋	●	●			●
📋			●		
📋			●		
📋				●	●
↪	■	📋	📋	📋	📋

Visualizing the control-flows

- On the bottom we can recognize the start activities, while on the right we can see the end activities.
- We can also observe the activities for which self-loops occur, namely the Assesment (553 times) and the Surgery (455).
- Further, note that the Handover emergency entrace is always immediately followed by surgery or resusciation.

Precedence matrix Example

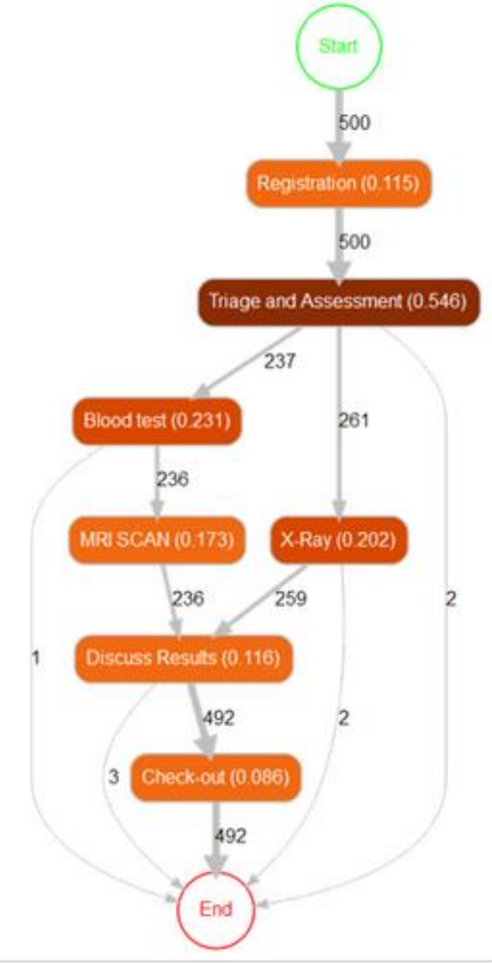


	Assessment	Check-In	Check-Out	Handover Emergency Entrance	Hospitalization	Resuscication	Surgery	Treat Minor Injuries	End
Treat Minor Injuries	108		257		120		235		
Surgery	251		391		206		455	112	
Resusciation	27						29	34	
Hospitalization									326
Handover Emergency Entrance						90	161		
Check-Out									648
Check-In	723								
Assessment	553						535	574	
Start		723		251					

Visualizing performance analysis

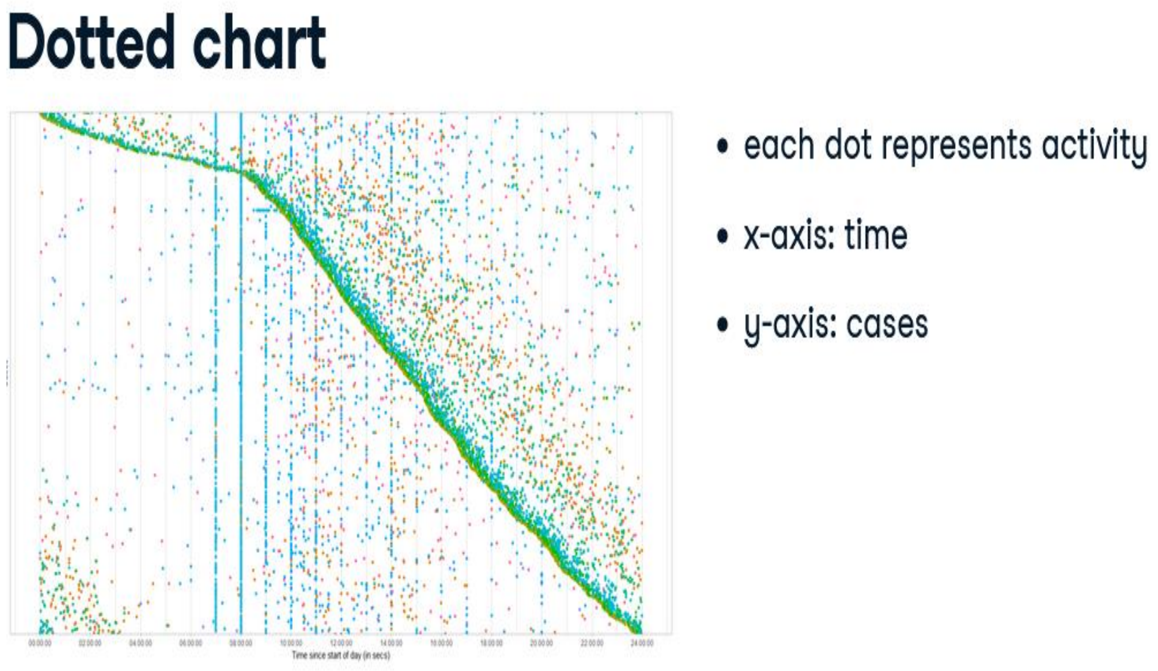
- The performance process map is a special type of process map which does not show frequencies on the arcs and nodes, but durations, both of activities and of the times between activities.
- The type is specified with the performance function, which can be further configured, for example by setting the aggregation that needs to be performed on the durations: mean, median, maximum, etc, as well as the preferred time unit.

Performance process map



Visualizing performance analysis

- Another specific technique related to time is the Dotted chart.
- While the performance process map focusses on the duration of activities, the dotted chart shows the distribution of activities over time.
- The dotted chart is essentially a scatterplot of activity instances each occurring at a specific time (x-axis) and belong to a specific case (y-axis).

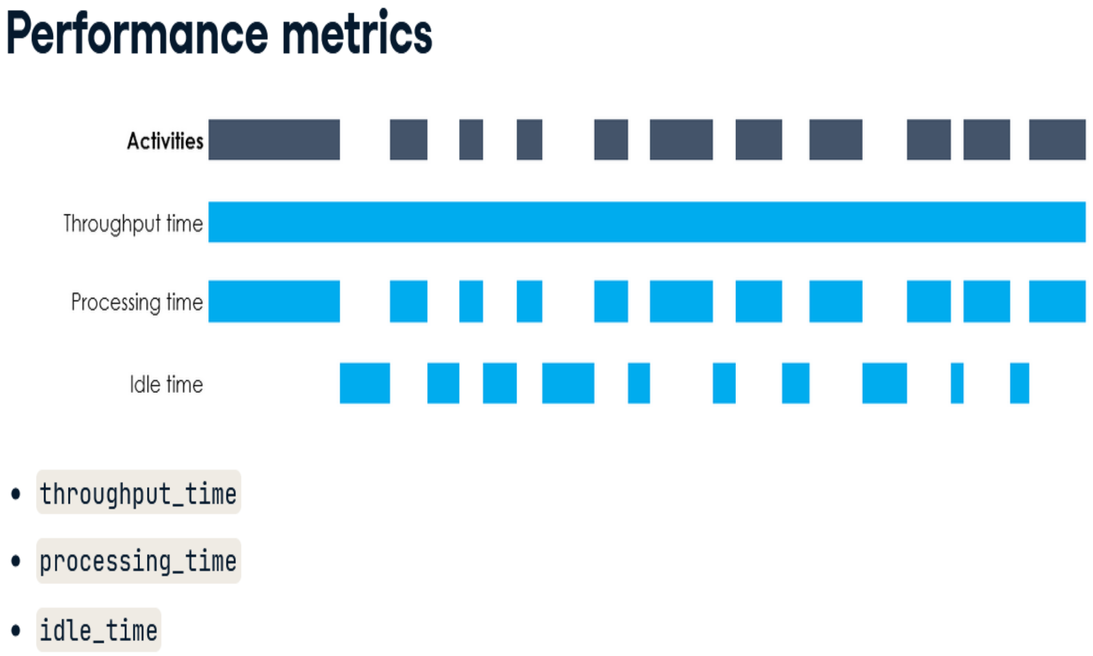


Dotted chart

- each dot represents activity
- x-axis: time
- y-axis: cases

Visualizing performance analysis

- A more numeric analysis of performance can be done using three different metrics, each expressing a different type of performance.
- Throughput time is the time since the start of a case until the end of the case, which include both active time and idle time.
- Processing time is the sum of the activity durations, which means it does not include the time in between different activities.
- Idle time is the sum of the durations between the activities, in which no processing of the case takes place.

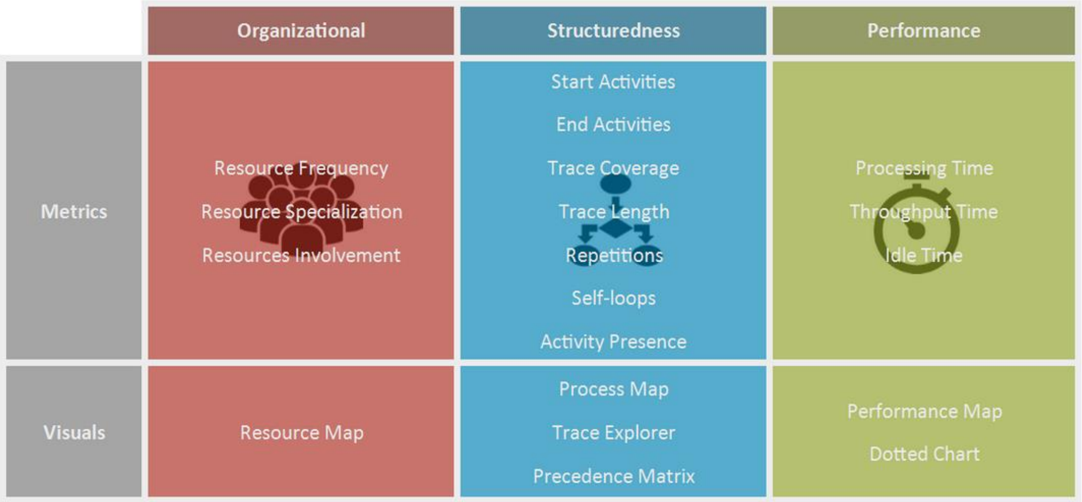


Performance metrics

- `throughput_time`
- `processing_time`
- `idle_time`

Linking different perspectives

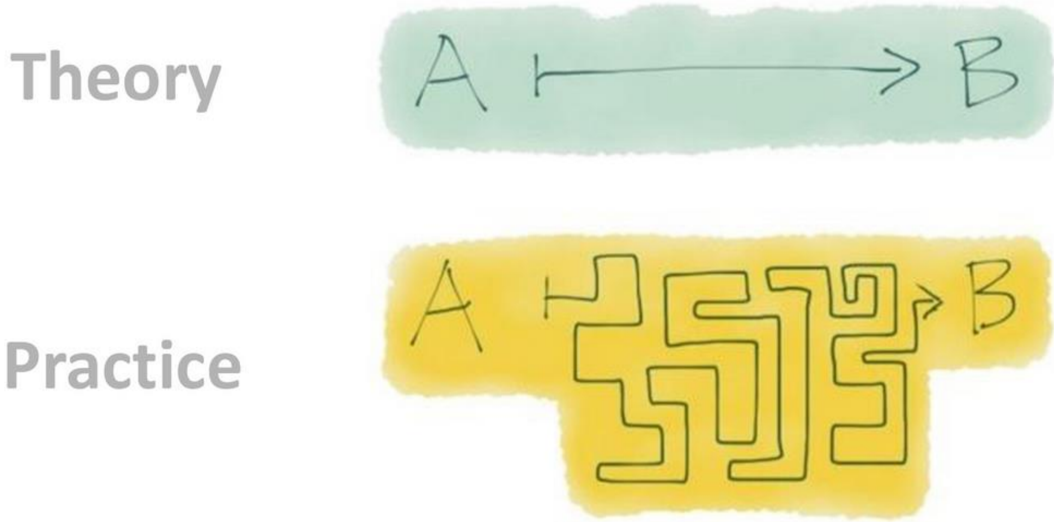
- Organizational
- Control flow
- Performance analysis



	Organizational	Structuredness	Performance
Metrics	Resource Frequency Resource Specialization Resources Involvement	Start Activities End Activities Trace Coverage Trace Length Repetitions Self-loops Activity Presence	Processing Time Throughput Time Idle Time
Visuals	Resource Map	Process Map Trace Explorer Precedence Matrix	Performance Map Dotted Chart

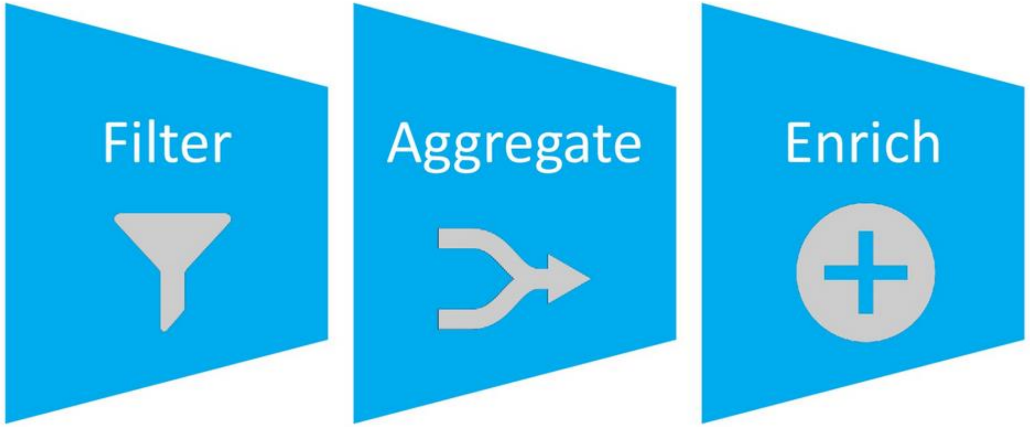
Filtering the data

- There might be too many activities, there might be cases of different types and time periods, which we do not want to look at simultaneously.
- Or it might be the case that certain data attributes are missing or not stored in the right format to be used in our analysis



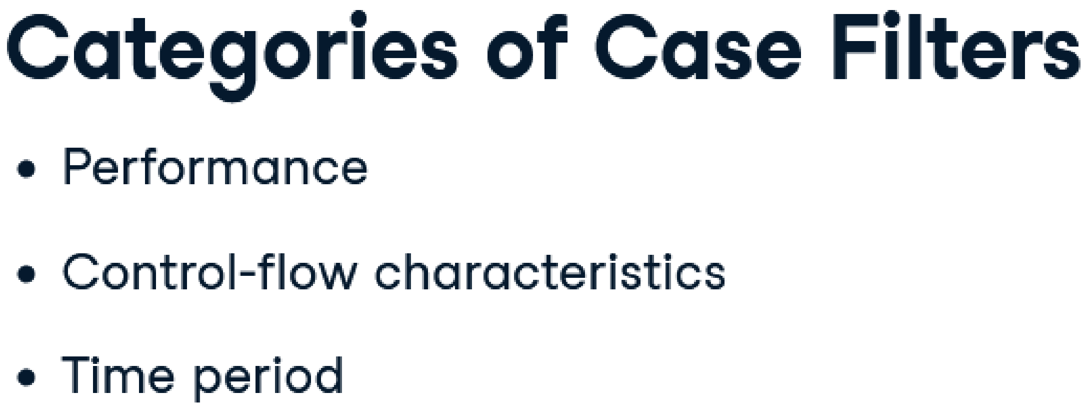
Filtering the data

- Data Preprocessing.
  - Filter
  - Aggregate
  - Enrich



Filtering the data

- From a high-level perspective, there are three categories of case filters:
  - cases with a specific performance,
  - cases with a specific control-flow characteristic,
  - and cases related to a specific time frame.

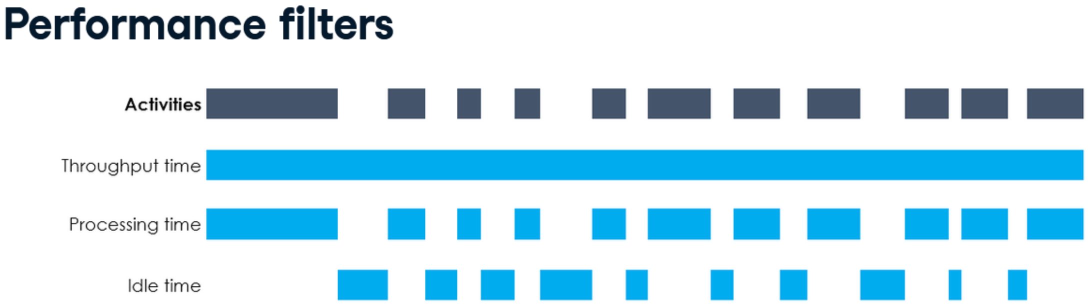


Categories of Case Filters

- Performance
- Control-flow characteristics
- Time period

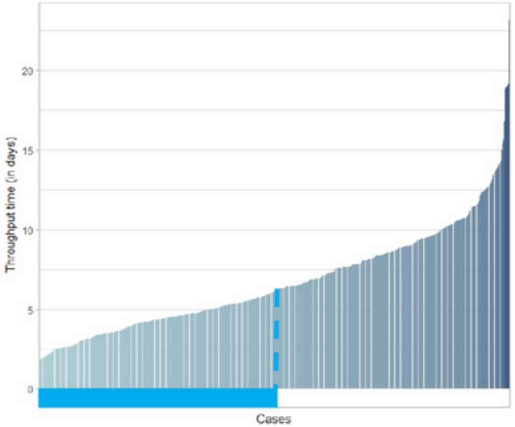
Performance filters

- We can consider four types of criteria:
  - the throughput time,
  - the processing time,
  - the idle time,
  - but also the trace length.
  - Performance filters are very useful to have a look at the long-lasting cases and check what went wrong, or to learn from the short, performant cases.

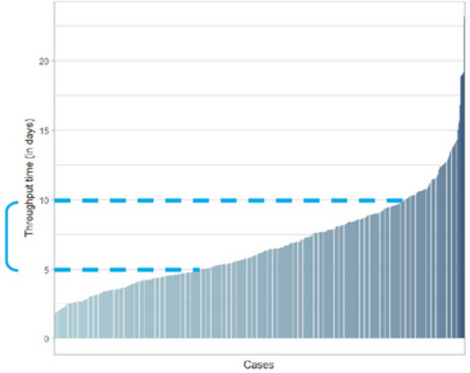


Performance filters

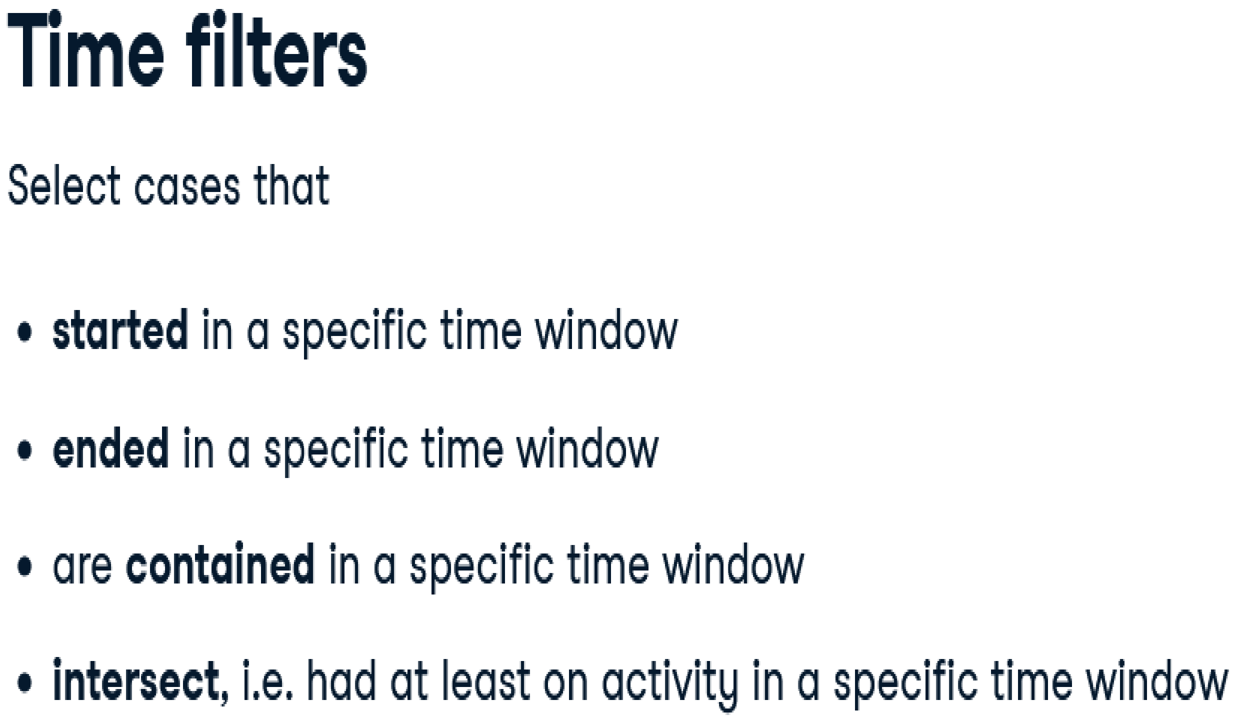
Filter by Relative Percentage



Filter by absolute interval



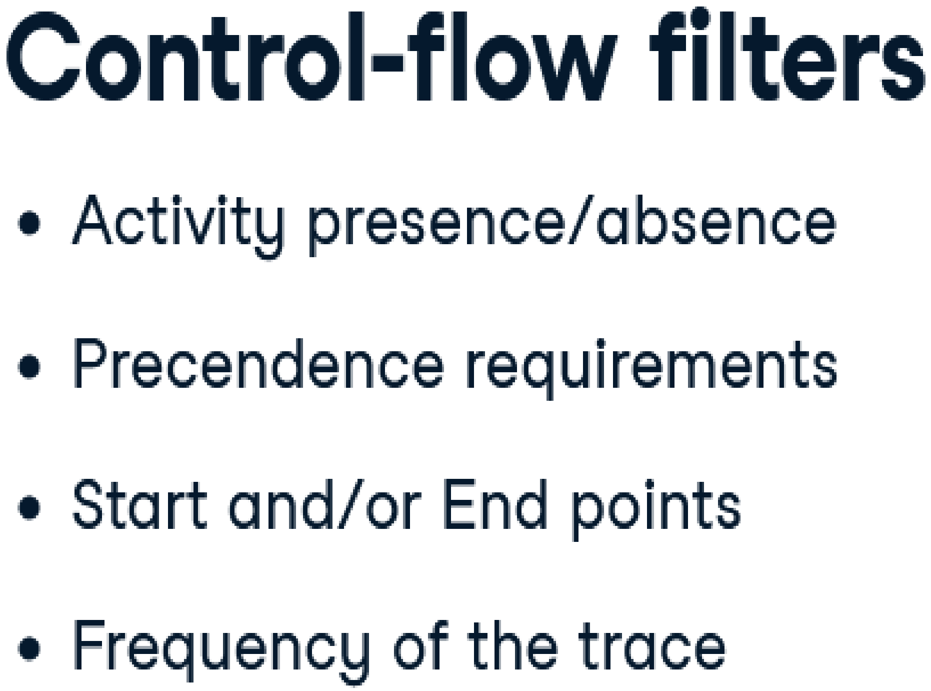
Control-flow filters



Time filters

Select cases that

- **started** in a specific time window
- **ended** in a specific time window
- are **contained** in a specific time window
- **intersect,** i.e. had at least on activity in a specific time window



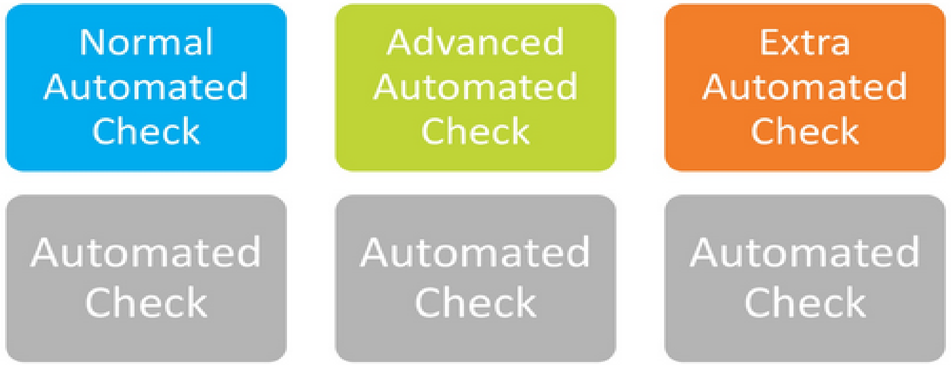
Control-flow filters

- Activity presence/absence
- Precendence requirements
- Start and/or End points
- Frequency of the trace

Types of aggregations

Aggregation types

Is-A



Normal Automated Check

Advanced Automated Check

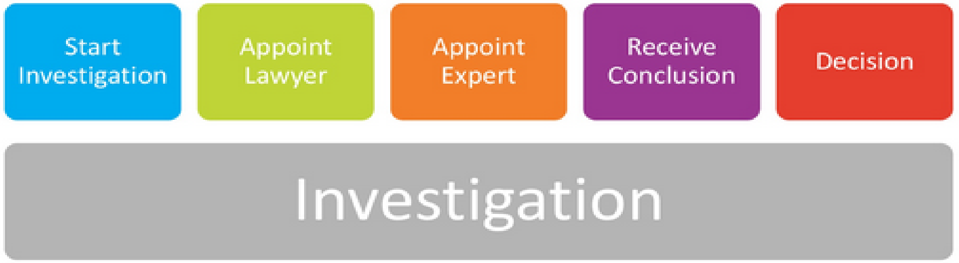
Extra Automated Check

Automated Check

Automated Check

Automated Check

Part-of



Start Investigation

Appoint Lawyer

Appoint Expert

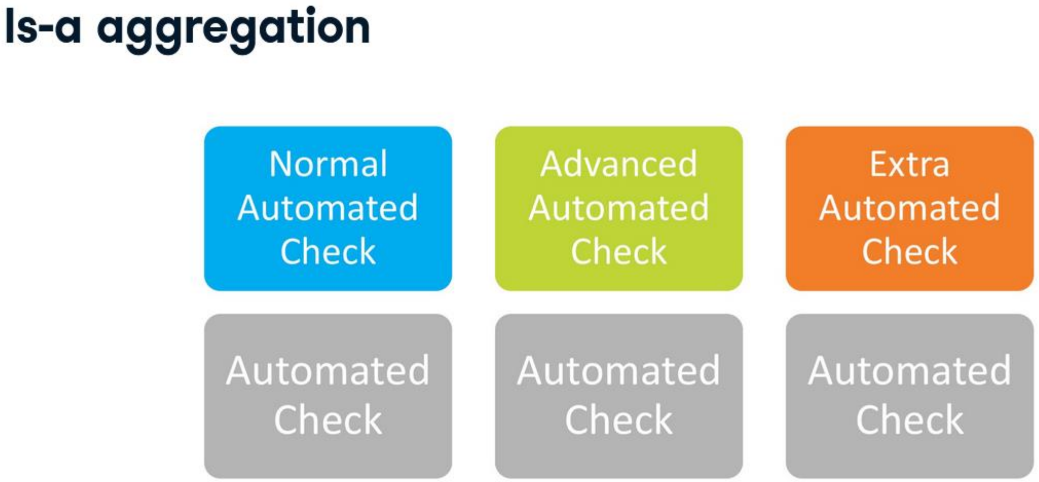
Receive Conclusion

Decision

Investigation

Is-a aggregation

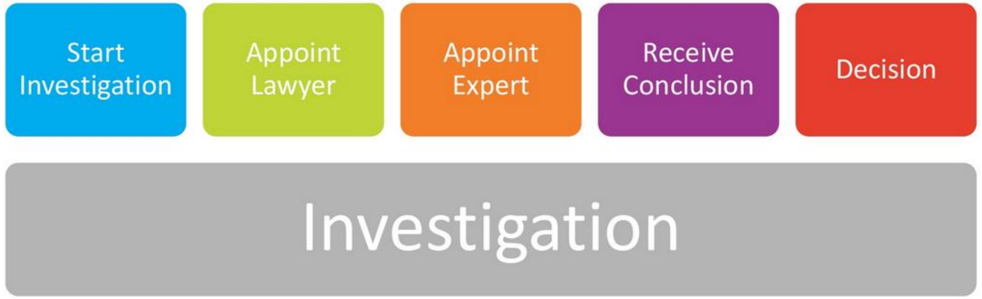
- The is-a aggregation occurs when there are slightly different activity labels, which at some level all related to the same activity.
- Suppose we are looking at an auditing process.
- We may have the activities “Normal automated check”, “Advanced Automated check”, and “Extra automated checks”. While the distinction might be useful for a certain analysis, suppose we are fairly comfortable with just calling these activities “Automated check”.
- This aggregation is called an is-a aggregation, because each of the detailed activities is a Automated check.
- Performing this aggregation can be done with the act_unite function, which means that we are about to unite several activities into a single all encompassing label.



Part-of aggregation

- On the other hand, the part-of aggregation occurs when there is a set of activities which are clearly different from each other, but they are nonetheless related as a part of a single, higher level activity.
- For example, in the claims management process, we have the activities “Start Investigation”,”Appoint Lawyer”, “Appoint Expert”,”Receive Conclusion” and “Decision”.
- While each of these is clearly a very different activity, they are all part of what we could call the “Investigation”.
- We can refer to this situation as a sub process, which we want to collapse.

Part-of aggregation



- Enriching event data is nothing more than adding calculated variables to the data.
- The traditional way of adding new variables to any dataset.

Enriching events