The role of Data in Business Processes

An edge through Data and Process Mining

Abstract: What are the different data components that Business Process Management Professionals hold at their disposal and how can the enterprise gain a maximum benefit from these? The present paper looks at the typical scenarios when it comes to harnessing the insights stemming from process data and to the new tendencies as well as technologies in that realm, specifically data mining and process mining.

Basic Information

When dealing with BPM (and RPA to some extend), business and process analysts typically face at least 2 important data pillars: transactional BPM based data and form & integration-based business data. This paper will go into the innards and usage scenarios of both sets, paving the way towards advanced analytic strategies, using machine learning, decision tree diagrams, sentiment analytics and more. Process users on one hand must gain a firm purchase on the run-time behavioral patterns, the in-depth analysis of historical data sets, and, on the other hand, have to increasingly combine existing information effectively in order to produce quality forecasts that then in turn feed into the process rules logic, adaptively. That's no small feat, especially considering that most users are still struggling with the most basic kinds of process data analytics: the review and interpretation of historical BPM data.

Figure 1 - Macro evolution of BPM Data; Winkler, Kay; 2019

Latest at the formal inclusion of dedicated toolsets for process analytics as a criterion to be considered an iBPMS vendor by the analysts, advanced forecasting that derive from business process data, has become a standard requirement by BPM users all over the globe.
Reports or – even more rudimentary – stored, raw process data could be described as the last piece of the "basic" elements of a BPM implementation. Now, while certainly advanced reporting, BAM, pattern recognition and predictive analytics are powerful features to accompany a process automation, covering the first elementary step of making sure that all the process as well as the business (form) data is stored in an automated, uniform, accumulative and (very important) scalable fashion – throughout all implemented business processes – is far more crucial (and more often than not something overlooked) for gaining real process insights and such enabling continued improvements. The key ingredients for viable business reports are in part derivatives of the process and form-variable design efforts and also in part the understanding of well-defined process metrics.

Of Process and Business Data

Having to deal with not only the statistical expressions of the process behavior but also its effects on the business end of things, certainly adds to the complexity a BPM analyst faces on a daily level but also – and more importantly – provides the users with the grand opportunity to gain an understanding of the causes and effects the process performance exercises on real life business outcomes. Add to the obvious advantages of continuously optimizing processes with these enhanced data sets the availability of increasingly more intuitive data science applications, it becomes clear why many experts in the industry confirm that BPM users are starting to take advantage of data science tool sets for process optimization to such a degree that analytics such as Data Mining have been adapted and coined as Process Mining in the field. And this for good reasons! On a per process bases, the process analyst can achieve quite the effect of scaled economies, given that with a relative few dependent and independent variables great insights on mission critical, end-to-end processes can be won (an example of such a variable will be detailed in a later paragraph).
As hinted on above, a typical BPMS produces (at least) two distinct sets of data that can be either attributed to information indicative to the elements of the business process as a such or to the data said process handles; be it through its forms, integrations to other processes or through integrations that feed and read data from and to outside applications.

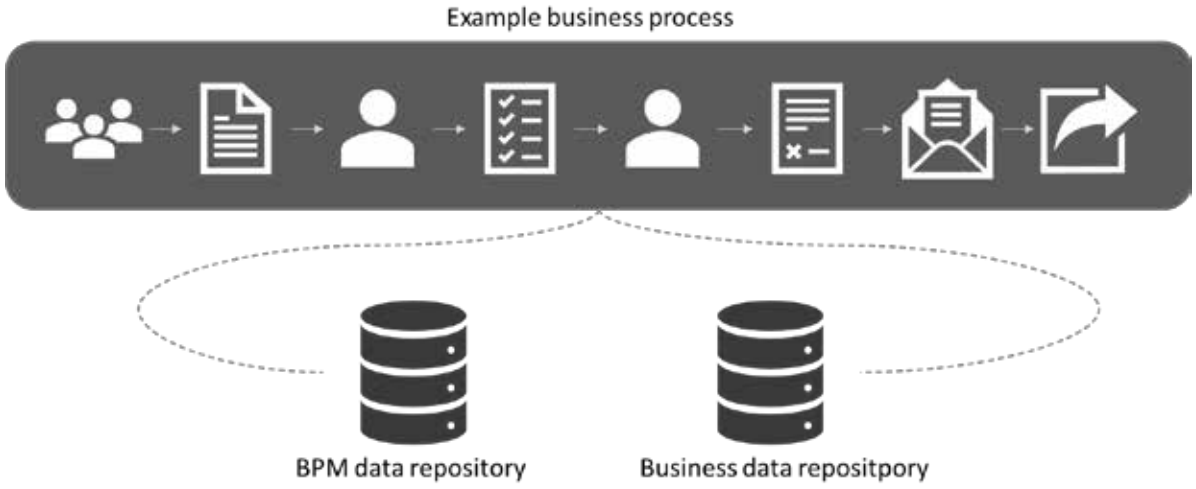

Figure 2 – Business and Process Data in BPM, Winkler, Kay; 2019

A typical example would be as follows:

For the BPM data repository (not in any specific order)

- ID (Unique process incident or case ID)
- Initial date (Date when the incident/case has been "launched")
- Initiator (User who initiated the incident/case; can be a named or anonymous user that can be later identified through other means)
- Concluded (Tag to identify if and how the incident has been concluded; for example: not concluded/still ongoing, concluded through case abortion etc.)
- Conclusion date (Date stamp if applies)
- Life cycle (Time stamp, if applies ofincident/case duration)

- Current process step (In which process step is the incident/case currently located?)
- Assignment date (When has been the incident/case been assigned it its present process step?)
- Assigned Users (To whom has the incident/case been assigned to?)
- Assigned Role (To which role/group does the assigned user belong to?)

Besides the essentials above, there of course can be many other, additional variables a given BPMS captures. It's important to notice that commonly this information is being "recorded" by most BPM platforms automatically and accumulatively, independent from whatever processes the engine handles. In that sense, one would have access to all the relevant BPM information regardless of the user's decision to use the BPM platform for automating a procurement or a vacation request process, for example.
With this information on its own a lot of knowledge can be gained. For instance, time per process and case can measured (more details on that below) and process based KPI's can be formulated (for example a goal of a vacation request that's not supposed to exceed 1 working day until decisioning and not longer than 2 working days until an approved request gets logged into the HR systems and notified to all involved users).

On the other one would have access to the wide variety of process dependent data which clearly depends on the specific business scenario the automation solution has been implemented for. This data can therefore be as diverse as the company's different processes, its forms as well as internal functions (business rules, calculations, integrations and so forth). In the vacation request process example from before for example, typical process specific data artifacts would be:

- Requester name, ID etc. (data likely to be declared in the form)
- Requester department (data likely to be declared in the form)
- Request approvers (likely the result of a business rule in conjunction with an integration to the active directory)
- Requested dates (declared data)
- Remainder of available vacation days (business rule with a calculation and likely an integration to the HR repositories)
- Decisions and reason codes (in case of rejections).

Both sets of information are of very indicative nature on their own but provide exponential insights when combined. Now, interesting correlations can be analyzed in the pursuit of continued operative improvements. Analysts will be able to detect, track and act upon patterns that stem from the influences that a specific user has on response times and case volumes, for example. In our example it would be of interest determining if vacation requests are being more likely to be reviewed late by a specific user and if these tendencies are altered by seasons, cycles or specific dates.
As a response, dynamic rules can then be implemented into the business process, provisioning alternative or additional resources if a given tendency breaches an established tolerance threshold.

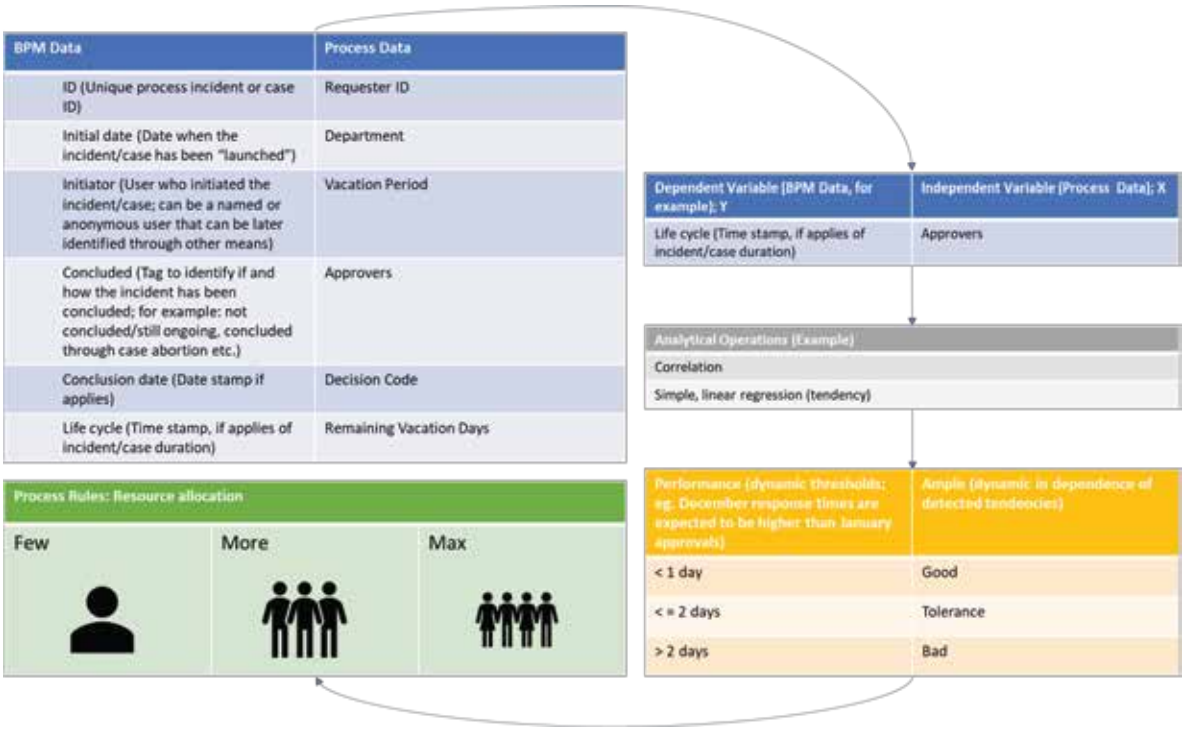

Figure 3 – Business and Process Dataanalytics example in BPM, Winkler, Kay; 2019

BPM Metrics, Time Variable

The Association of Business Process Professional's Business Process Management Book of Knowledge (ABPMP CBOK) lists the process performance measurement as a fundamental element of the BPM practice (ABPMP, 2009). Depending on different possible viewpoints within and around the process environment, those measurements may form the foundation to justify a BPM engagement in the first place (project evaluation), identify the ROI of such an engagement – post project or to pursue continued improvements to existent processes.

Establishing viable metrics can at times entail severe challenges. Trying to establish a robust framework of measurable variables that encompass the most crucial business process patterns without causing an excessive accumulation of sometimes redundant data points typically requires the design team and the process owners alike to critically boil down the most representative dependent and independent key variables to be measured throughout the company's processes. This definition of measurement variables in turn should then occur ideally early in the process design phase and take into consideration challenges like confidential information (example: hourly costs of employees for human centric processes) that likely won´t be captured on the process level.

One of the most common and intuitive metrics that BPM users, analysts and providers refer to is the process cost measurement. In praxis however, especially in human centric processes, sometimes due to the confidentiality of resource cost information and sometimes because of its unavailability at design time (pointing to whole different set of additional challenges), process costs (ex-ante and

ex-post BPM) often can't be exactly and easily represented as monetary values (leading to sub-sequential and colorful ROI guesswork).

An alternate and eventually more generalized approach to that dilemma maybe the declaration of "TIME" as the principal and dependent core variable for all economic process measurements, which of course will suffice only for service processes whose material inputs play a merely secondary role to its results. Different types of "TIME" can easily and natively be captured by most BPM platforms and later be analyzed and compared within the same company (different processes and versions) and be benchmarked among different corporations. Different economic scenarios, tactics and strategies can then afterwards provide monetary multipliers to the continuously measured "TIME" variable, delivering a dynamic business context to process owners and analysts, on demand without being process costs an embedded BPM metric.

The core challenge hereto would be the definition and the coherent measurement of different types of "TIME" variables. One can differentiate (among other types) between the task lifecycle time in each process and the worked time of different individuals for a task during its lifetime. Drilling down further, one could argue to differentiate the overall task lifecycle time and only its workday lifecycle time (net lifecycle) and so forth.

Having established "TIME" as a dependent core variable for process measurements and the possible formulation of behavioralhypotheses, result influential independent variables can now be included into the process metric framework.

As detailed later on we have found that for most human centric business processes "net task lifecycle time" as a dependent variable and "quantityof process handoffs", "quantity of monthly process transactions" and "quantity of fully automated core and legacy integrations"– all as independent variables– have shown a significant correlation to each other as a result of BPM engagements.(Winkler, 2013)

As summarized previously, thereare several possible definitions of time in a process that can be categorized as:

Process extension time: The sum of all configured extension times of a process' steps. Typically, during process design time, the process owners define how much time each process step should take. This definition can be different from the task time and worked time.

- Includes Step Extension Times

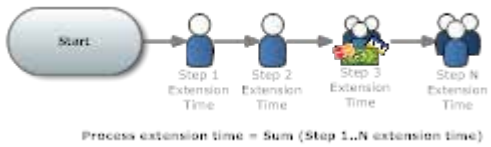

Figure 4 - Process Extension Time, Winkler, Kay; 2019

Process worked time: The sum of all configured worked times of a process' steps. Typically, during process design time, the process owners define how much working time each

process step should consume. This definition can be different from the task time and extension time.

- Includes Step Worked Times

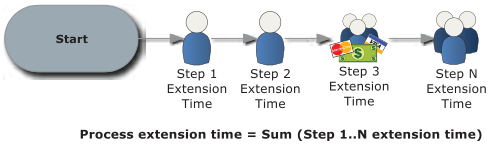

Figure 5 - Process Worked Time, Winkler, Kay; 2019

Task cycle time: The entire lifecycle of a given task within a process, summing up all cycle times (calendar time or only worked time) of that task within the different process steps until it comes to a cancelation or "natural" conclusion of that task.

- The task time takes also into consideration task redundancy cycles due to errors or policy violations (reworking tasks in previous steps).

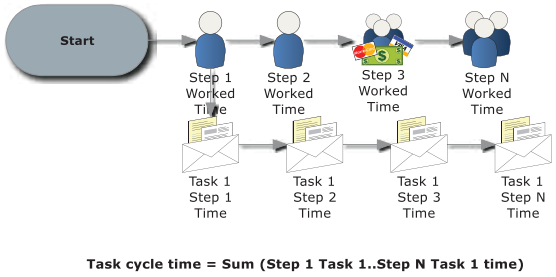

Figure 6 - Task Cycle Time, Winkler, Kay; 2019

While "Process Time" is something (as shown in the BPM methodology section) that is defined during design time and is the rigid instruction of the process to be (time wise enforcement of a worst case scenario that is ensured by a BPMS escalation and notification system), the "Task Cycle Time" would be its dynamic counterpart of the actual process time consumption during runtime. Hence "Process Time" could be classified as a designed metric and the "Task Cycle Time" as a measured metric. In relation to each other: "Process Extension Time" > "Process Worked Time" > "Task Cycle Time". The here presented results correspond to the "Process Worked Times" (the pre-defined process times during its design phase).

Measuring time is measuring the most important economical variable of human centric processes that then can be freely multiplied with whatever hourly rate and real-life scenario that may apply.

For all time types above, the final sum of each expresses a single process run (from "start" to "stop" of the process).

Evolution towards Deep Learning

Crawling before running is also good advice when it comes process performances measuring (PPM). While it is true that PPM itself is becoming increasingly accessible for business and none technical end users, a couple of basic components must be put in place first, before the desired fruits from machine learning (ML) or artificial intelligence (AI) applied to the process world can be reaped. As pointed out above, there is a great wealth of information to be understood and effectively used for process enhancements, coming from the BPM platform alone. Enhancing said data sets with business data that is fed into a unified data mart or dedicated repository, opens the doors for broader correlational analytics through the means, for example, of time series or cross-sectional investigations.

Besides architectural recommendations like running query laden studies of process and business data against a database that is separated from the BPM production environment and dedicated to these means, there are also other important design decisions the business analysts have to take. Many experts in the field recommend establishing strategic measurement goals first, making sure that all required variables are available and complete. A more recent school of thought emerges from the field of Big Data, where the assumption states that basically all existing information is being stored, made available and where during the analysis the pursuit of correlation patterns is of more importance then a pre-established search and proof of causality. In both cases however, the integral traceability of transactional data that spans all the different systems that are involved in a business process is of utmost importance. Also, when it comes to system crossing data it can be initially counterintuitive to not limit the data-trace and analytics on the BPMS repositories alone, especially in the case of process mining. There, hash value anchors as well as modern blockchain technologies can be of great assistance for connecting the dots not only between different systems but even between different organizations altogether.

Gaining meaningful and deep insights from business processes can be, understandably, a daunting undertaking, should therefore organized in stages:

Given a sound system architecture, healthy and integral, transactional and process data repositories, stage 1 would represent the efforts of understanding the trends and possible patterns of the basic process main variables ("time" for instance, as detailed earlier). Applying (lineal, quadratic, simple, multiple…) regression techniques each process variable should be first analyzed individually and later correlated with other process variables, testing different models for (positive or negative) correlations and predictive qualities. The process response-time variable, for example, could be reviewed on its own, identifying through a time series if specific trends or trend patterns exist. Assuming several years' worth of daily process case registries, indicative observations regarding response time trends, cycles and seasonality can be made. These "vectoral" understandings can be broadened by amplifying process data models with additional BPM native variables. For instance, the business analyst can put case response times in correlation with process complexities, expressed by the quantity of process steps (hand-overs), through a cross-sectional analysis. Alternatively, a specific variable can be closer inspected after having determined certain

trends, in order to identifying causality.The following example showshow the simple counting of active incident (case) quantities per process step over time, could provide cluesto the overall workload or could help explain response time patterns (if observed monthly):

Figure 7 – Net Production; Winkler, Kay; 2019

In this hypothetical caseof a lending process, incidents are listed per process stepand grouped into statuses "initiated", "carry-over" (from the previous month), "completed" and "aborted" (in real life many additional incident statuses, depending on the BPMS, can apply). The "Net Production" portion of the matrix then details to which originating month the completed production pertains. With such a simple visual arrangement, the answer to thepossible question for more resources to face slower response times and bottlenecks can be refined.

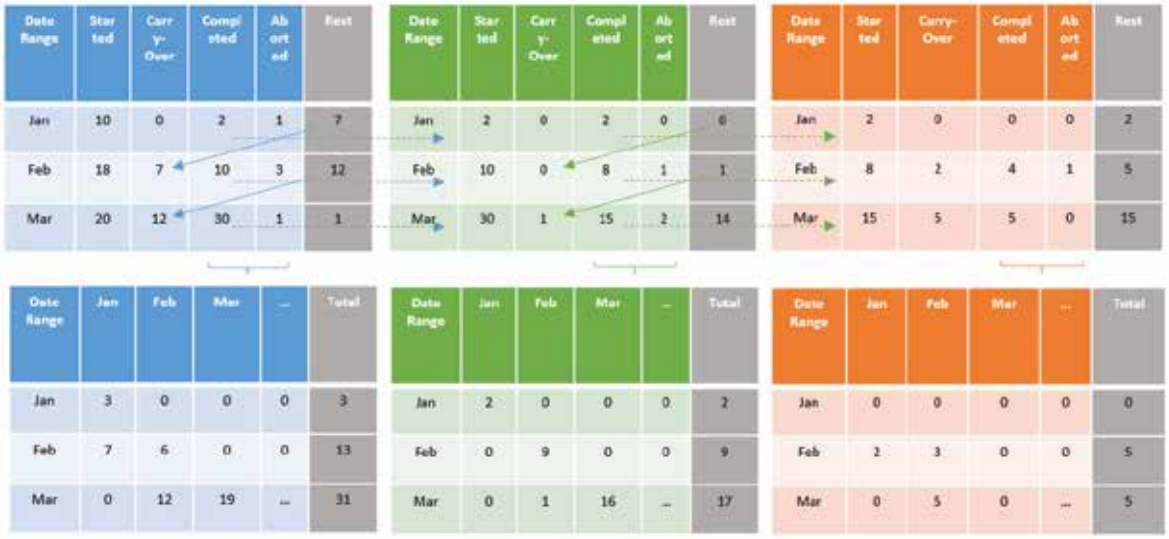

Figure 8 – Net Production amplified; Winkler, Kay; 2019

The continuing stage of data analytics in process could be divided into the data visualization phase and basic pattern recognition, as the following or parallel step.

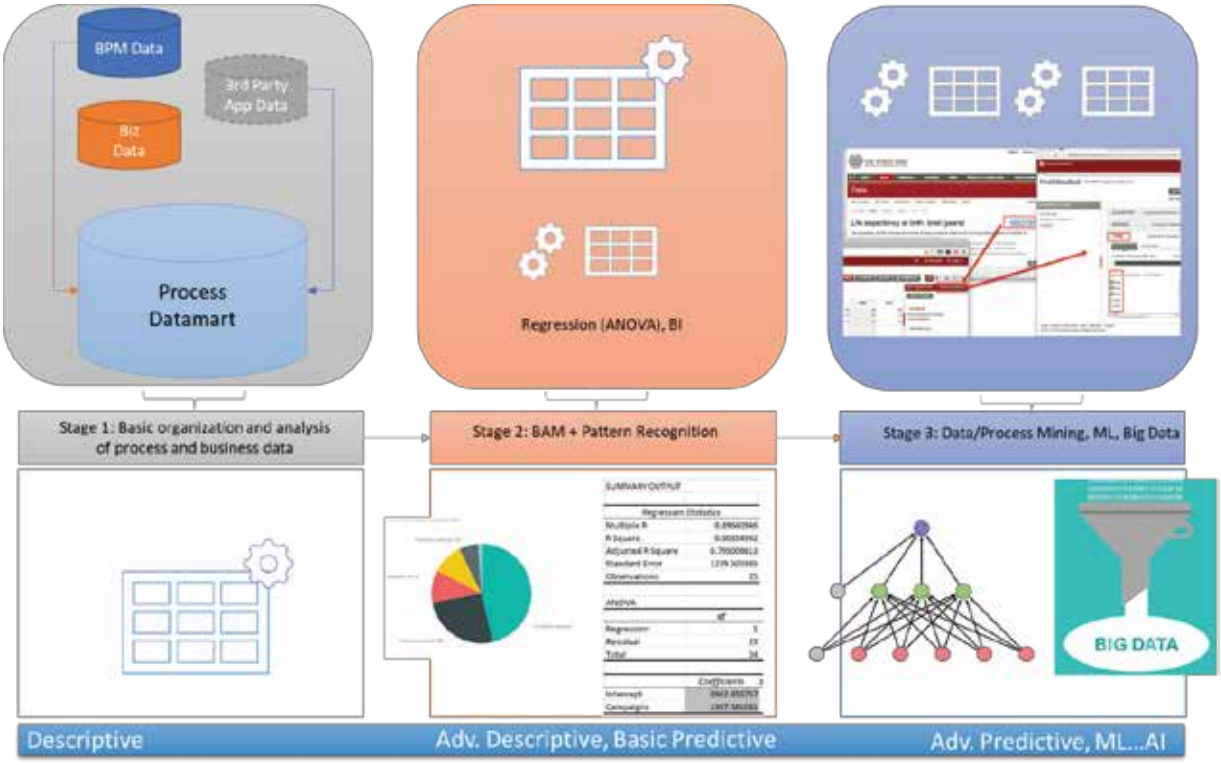

Figure 9 – Stages of Process Analytics, Winkler, Kay; 2019

Typically, the stages of process analytics are not exclusive but rather interdependent and complementary to each other. Also, depending on budgetary constraints and specific business goals, developing a domain expertise within a specific stage may be more suitable for some companies than by default having to achieve "ML level analytics" all the time as the efforts considerably grow, the further right one evolves within the process analytics spectrum.

Advanced Process Analytics

In addition to simple data visualization activities, like the review of the monthly net production resulting from a specific process shown previously, there numerous other activities the business analyst can undertake from the base BPM and business data repositories:
One such (recommendable) analysis would be, of course, the review of process response times over time to begin with. Here, the evolution of response times could be measured through monthly observations and then put through a time series analysis, validating for trends and statistically indicative forecast models. A study from 2014 found (Winkler, Benefits of Policy and Rules Driven Processes in LatAm Retail Banking Automation, 2014) that a newly automated process in the retail banking industry, for instance, typically accounts for approximately 35% of savings in worked time, when compared to the none automated process models. The application of a time series application can then further these statistics over a longer period and also start looking at seasonal impacts that may occur.

Time Saving Scales (0 - 400 working hours)	Quantity
Range A -0-9.99	32
Range B -10-49.99	38
Range C -50-99.99	31
Range D -100-199.99	17
Range E -200-399.99	5
Grand Total	**123**

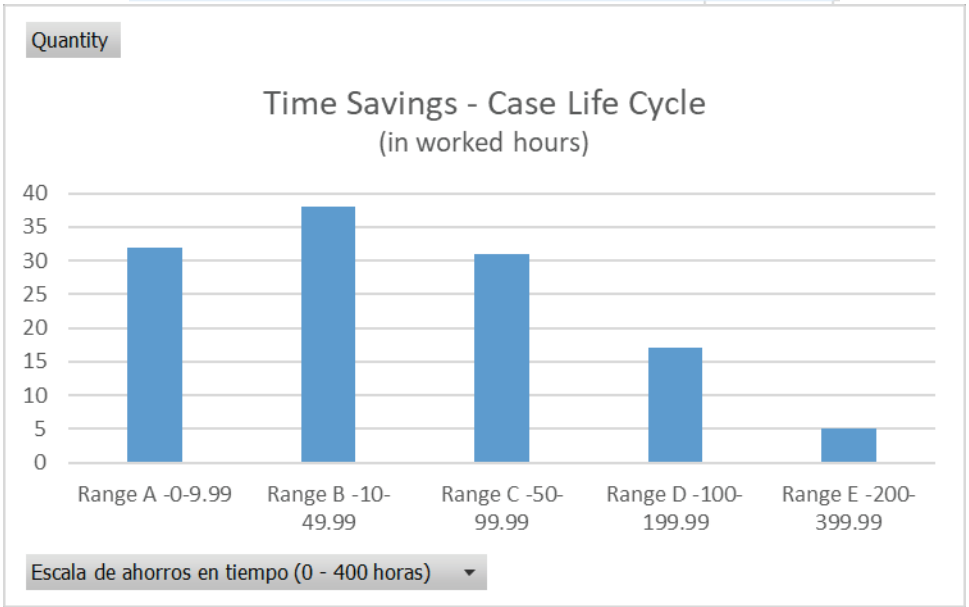

Figure 10 – Life Cycle Time Savings of BPM, Winkler, Kay; 2019

In business process management good data science practices apply as well. For example, data integrity and completeness are at least as important as a "healthy" amount of quality data points. It just wouldn't be worthwhile trying to establish a tendency of improving response times within a commercial sales process, when only basing the initial regressions on some couple of dozen monthly observations.

That's typically when the companies' BPM and data science maturity comes into play, allowing a certain period of investments to take root before being able to gain the first returns on investment. Having achieved this level of insights however, furthering investigations will become far more accessible and easier to implement. With a complete and well-structured process repository, modern business intelligence platforms can easily be used as bridges to extend the BPM data sets to the business data and with that offer correlational links and provide visual aids of otherwise complex interdependencies and relationships

Year	Quarter	Month	Count of Incidente
2017	Qtr 3	September	2
2017	Qtr 4	October	23
2017	Qtr 4	November	14
2017	Qtr 4	December	14
2018	Qtr 1	January	23
2018	Qtr 1	February	13
2018	Qtr 1	March	10
2018	Qtr 2	April	10
2018	Qtr 2	May	16
2018	Qtr 2	June	15
2018	Qtr 3	July	2
2018	Qtr 3	August	6
2018	Qtr 3	September	10
2018	Qtr 4	October	12
2018	Qtr 4	November	7
2018	Qtr 4	December	2
Total			**179**

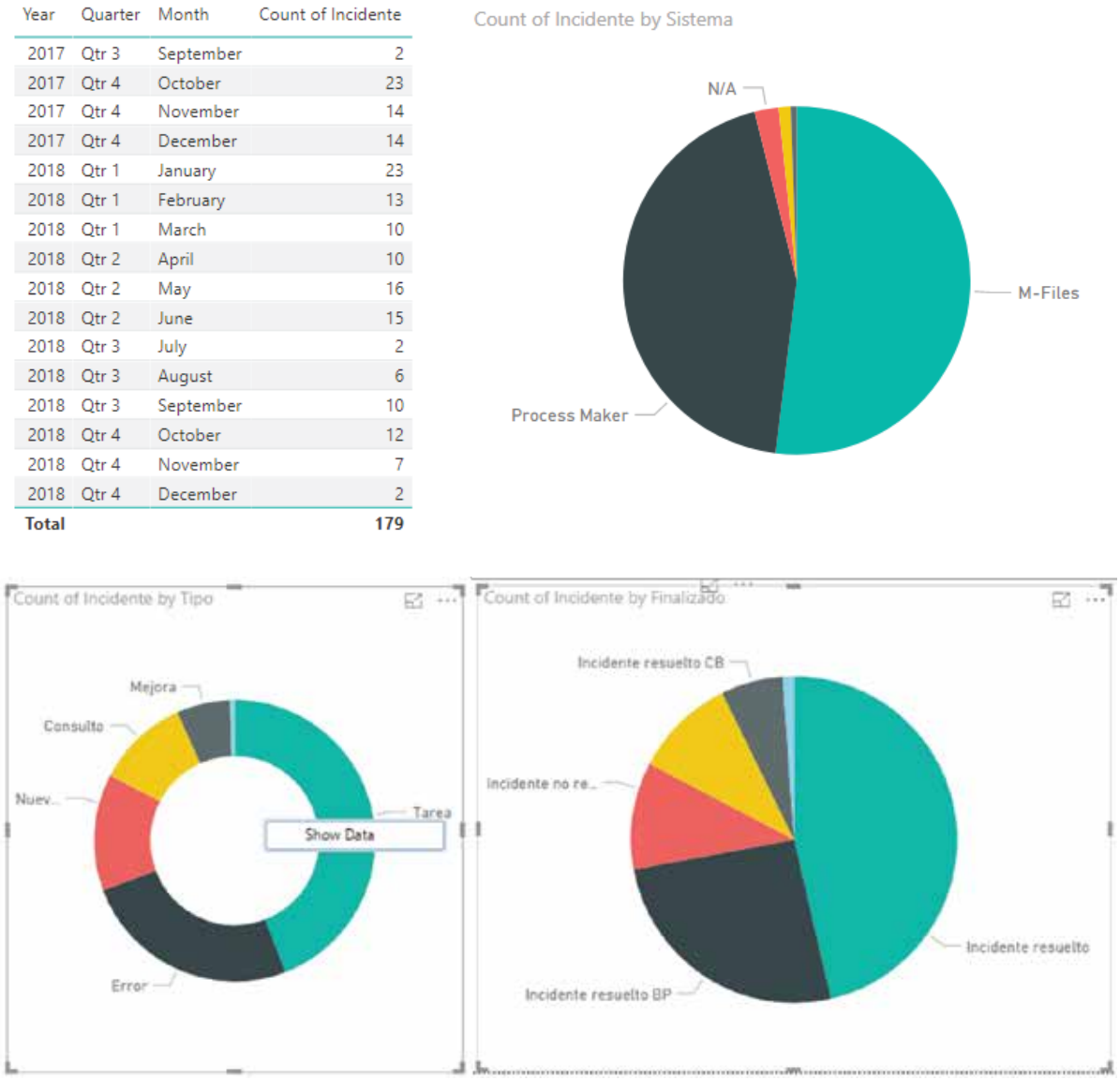

Figure 11 – Support BPM processdata presented in a BI Dashboard; Winkler, Kay; 2019

In the example above, there are different process KPI's that are easily displayed in a unifying dashboard that uses the data output of a process within a BI platform. The resulting console visualizes the concentration of processed cases per request type, user ID, and lifecycle extensions. In that sense, an efficiently organized representation of indicators relevant to the business strategy is ought to be designed in way that helps establishing and analyzing the relationships of the most result influencing variables of a business process. What kind of request type, for instance, accounts for the highest cycle times and is stemmingfrom which department and which user?
It becomes clear that such a Business Activity Monitor is the natural precursor to a deeper, relational analysis.
Besides the amazing diversity of different technology solutions here, this specific type of data reviews almost all the time bears just simple regression statistics at its core.

The aforementioned time-series, establishing time as the independent variable and any other given BPM or business variable as its dependent counterpart, allows, forexample, the initial review of process behavior over time and resulting patterns.
Considering several variables at once, permits the business analyst to create more complex data models and the not immediately obvious influencers of process (hence business) outcomes can be contrasted as a result.
In the referenced study, cross sectional analytics were applied to detect the influences the quantity of handovers and third-party application integrations in process have over time savings and paper consumptions. So, this study showed that an average of 35% of savings on worked time per observed process has been achieved, resulting in about 55 less working hours per case life cycle. More interestingly however, the correlational study of the different reviewed BPM variables showed that the level of the process as-is documentation as well as the quantity of user managed process policies had the most influence over these savings.

Things will become even more thought-provoking when extending time series and cross-sectional analytics to the fields of big data and machine learning for BPM.
Applying the Hodrick-Prescott filter, for example, to time series such as GDP and process response times in the BI software of your choice, correlations can be researched and categorized in terms of their quality as leading or lagging indicators for business outcomes through process management. A growth in gross domestic production may have a (x number of months) delayed growth impact of process volumes, leading to bottlenecks and these in turn into slow response time trends.
The sheer amount of the increasing number of publicly available data sets allows for an equally large number of investigations that can be undertaken in search for formerly unknown relationships of external "stimuli" to the business innards.
As indicated by Cukier and Mayer-Schoenberger in their piece "The Rise of Big Data" in the 2013 issue of Foreign Affairs, analytics in Big Data favors correlation over causality, serving also as quasi antidote of the McNamara fallacy when amplified sufficiently. (Mayer-Schoenberger, 2013)
On the opposite spectrum of Big Data, the discipline of data mining has made many important strides, to a point of creating its own niche in BPM in the form of Process Mining. Pioneered by thought leaders in the field, with Wil Van der Aalst at the very forefront, Process Mining has established itself as an important field of investigation among business analysts.
As initial step of the mining process, using unique transactional identifiers, flow sequences can be tracked within a given BPMS environment, as shown in the following scenario within Bonitasoft' s BPM platform, leveraging a custom Process Mining module.

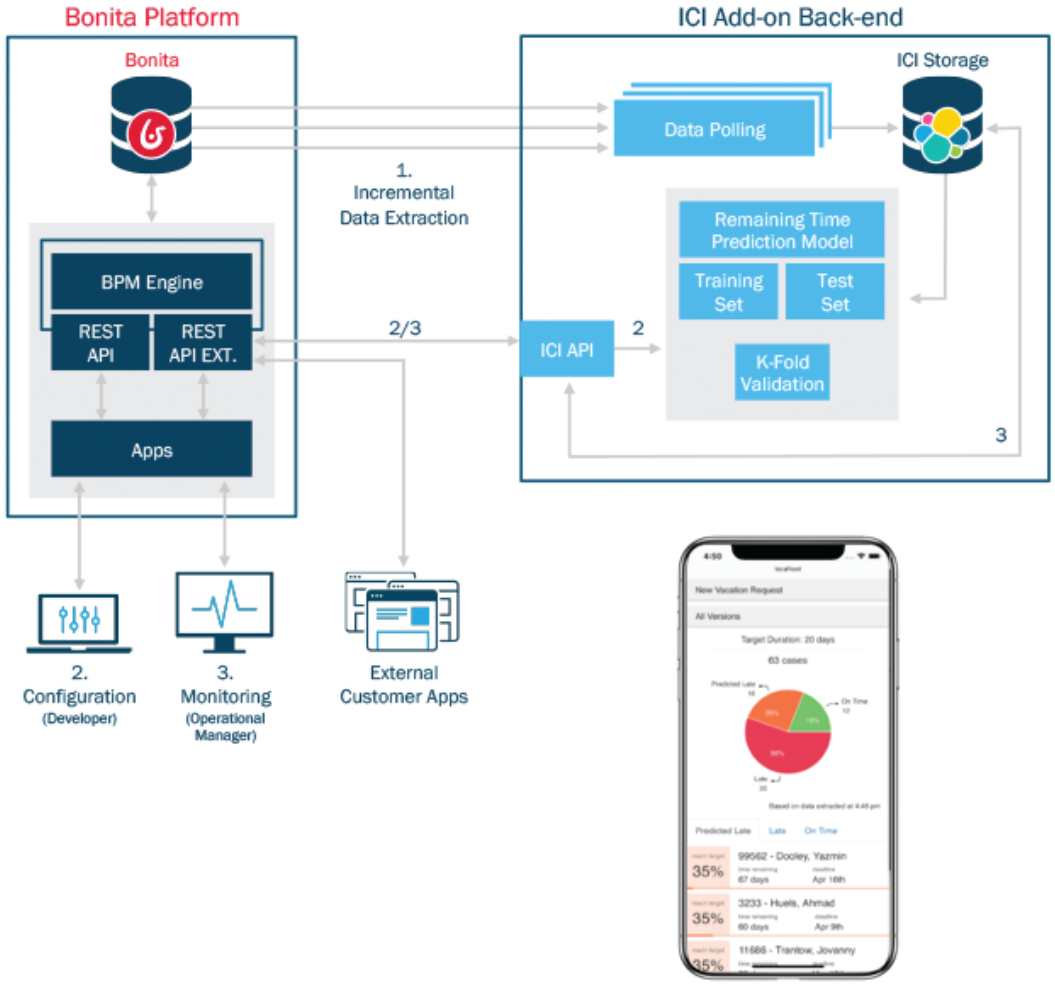

Figure 12 – Intelligent Continuous Improvement-ICI; Bonitasoft; 2018

Automatically tracking thereal-life workflow sequence is especially helpful to identify an accurate representation of the process' as-is situation as well to track the delta of how the sequence is ought to progress vs. reality.
Outliers can be caught earlier, its impacts studied, and processes be adjusted more efficiently aligned to the true operational workings of the company.
When during the data and architecturaldesign phasesthe unique transaction IDs have been created in way that integrity can be assured even when a case has officially left its processes of the BPMS environment, further, multi system and - with Blockchain technologies- even multi company tracking and process mining can be accomplished.
Process Mining introduces entropy calculationsas a measure for disorder, enabling the business analyst to visualize workflow patternsin terms of their homogenyinto graphical decision trees
With that in mind, following individual cases down the entire line of company's value chain, the impacts process deviations within aback-office workflow have on the overall customer churn rate, will become visible and, above all actionable.

In the paper "Decision Mining in ProM" the authors point out that "in order to analyze the choices that were made in past process executions, its required to find out which alternative branch was taken by a certain process instance. Therefore, the set of possible decisions must be described with respect to the event log. Starting from the identification of a choice construct in the process model a decision can be detected if the execution of an activity in the respective alternative branch of the model has been observed, which requires a mapping from that activity to its "occurrence footprint" in the event log." (Rozinat A., 2006)
If decision or process path mining, Machine Learning offers completely new and additional insights in BPM that many software vendors facilitate their users at the easy of just a couple of clicks.

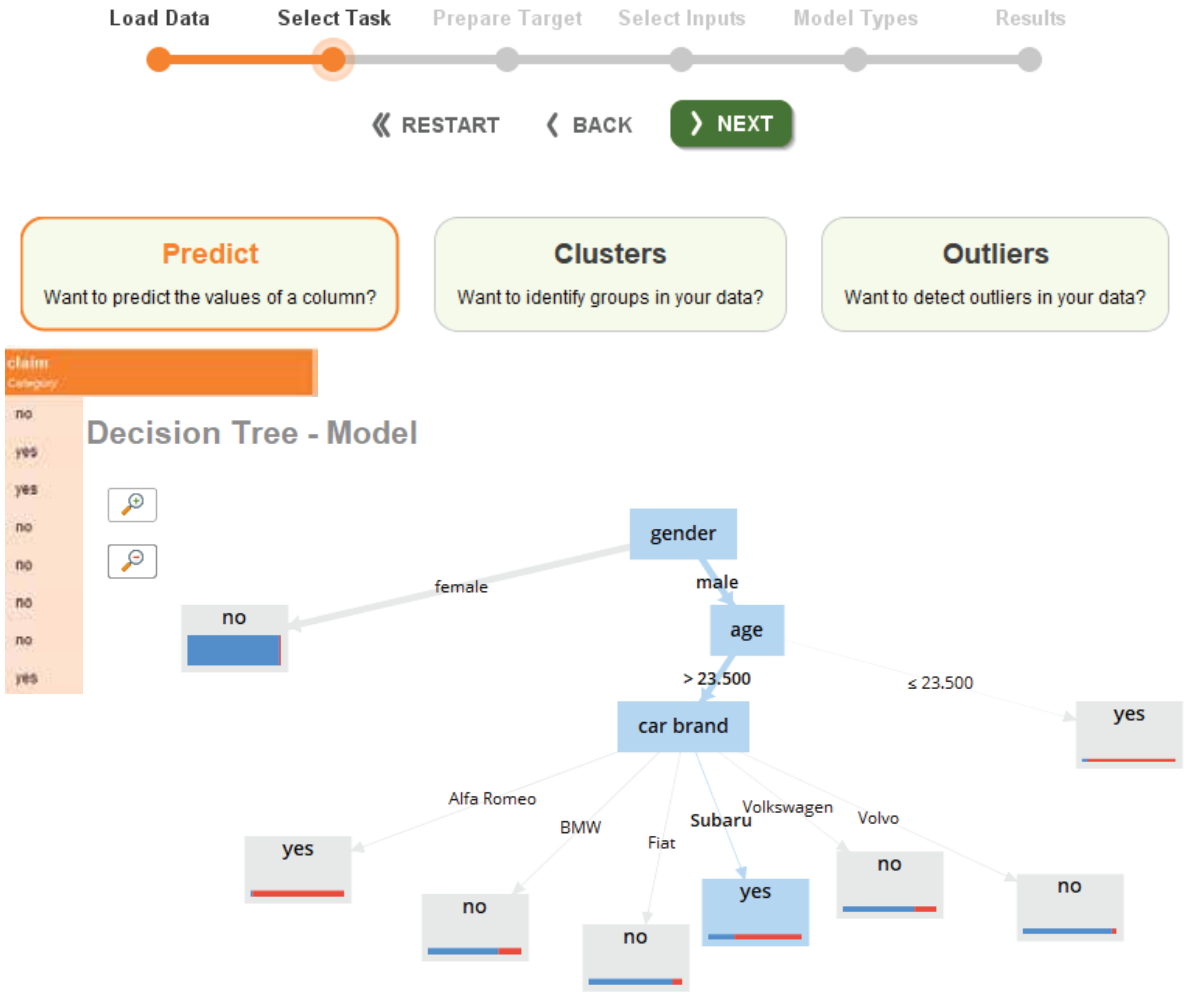

Figure 13 – Decision Tree Analysis with RapidMiner, 2019

The previous graphic shows how modern analytic platforms (in this case RapidMiner) have been become accessible enough for the everyday executive to run complex queries and operations against existing databases and gain a meaningful understanding with relatively little effort (in this stock example the decision tree of different predictors leading to insurance claims) Harnessing the power of these features, of course keeps relaying on the prerequisites of a sound technical

architecture, a cultural business mindset favoring data scienceand also the availabilityas well as the access to the data itself.

References

Mayer-Schoenberger, K. C. (2013). The Rise of Big Data.Foreign Affairs, págs. 28-40.

Rozinat A., v. d. (2006). Decision Mining in ProM. Computer Science, vol 4102

Winkler, K. (2013, February). BPM Leader. Retrieved March 22, 2013, from Viable Metrics to Justify a BPM Project: http://www.bpmleader.com/2013/02/11/viable-metrics-to-justify-a-bpm-project/

Winkler, K. (2014). Benefits of Policy and Rules Driven Processes in LatAm Retail Banking Automation. En L. Fischer, iBPMS: Digital Edition. Miami: Future Strategies Inc.

Figures