Subjective Perception Scoring

Psychological interpretation of network usage metrics in order to predict user satisfaction

Jörg Niemöller
Business Unit Support Solutions
Ericsson
Stockholm, Sweden
joerg.niemoeller@ericsson.com

Nina Washington
Service Systems Research
Ericsson Research
Stockholm, Sweden
nina.washington@ericsson.com

***Abstract*—Experiences of users determine their actions and therefore have great influence on the business result of service providers. Thus, it is highly important to know and control these experiences. Decisions and actions are most effective if they are based on individual users and their detailed experience history. Currently available methods for measuring user experience fail in providing sufficient detail or continuous availability. This paper introduces a big data analytics method, which predicts every user's satisfaction with the service provider. This is performed using an analytics algorithm, adding psychological interpretation to the measured and individually perceived user's quality of experience. The result is a score that enables individualized marketing and allows understanding the cause of dissatisfaction. It is also able to assign a subjective quality score to network assets.**

Keywords—Data Analytics; User Satisfaction; Quality of Experience

I. INTRODUCTION

The "segment of one" is a concept describing a trend in customer experience management to treat every customer individually. This implies that a customer's individual experiences, needs and behaviors are understood in detail for optimal decisions and actions. The main goal is to raise the quality of experience for each individual customer, with optimal usage of available budget and resources. The result can be high retention rates, reaching favorable recommendations and potentially stimulate more service usage. It is the fight for every customer's loyalty in saturated competitive markets.

It is only possible to manage, what can be measured. This paper introduces a big data analytics method that predicts a satisfaction score for every user. The score is called Service Level Index (SLI). If, at any moment in time, a user would be asked how satisfied he or she is with the service provider, this score would predict the individual answer of the asked user. This means that knowledge about individual satisfaction is available for every user and anytime.

The service level index is directly supporting individualized marketing. It can for example be used as a selection criterion for users who shall be included in a campaign for up-sell or remedial actions after an experience degrading incident. An individual detailed quantification of individual momentary satisfaction can be an essential part of the decision logic. The service level index introduces a new level of detailed control to these use cases. It allows scaling the effort put into remedial actions for optimal use of resources and budget.

Another use case would be a perception score of network assets. The idea is to assign a score to nodes and services in order to rate their quality not only according to objective technical quality thresholds, but by considering how a network asset's performance is perceived by the user. One example is the performance of radio cells. The metrics about the cell might be well within good ranges indicating decent performance, while the users, who are served by the cell, still perceive the service quality as bad. The proposed perception scoring model allows understanding and controlling this deviation between technical performance and subjective individual perception.

This paper introduces the Service level index as a big data analytics algorithm that is directly applying insights from psychology in order to interpret the data and generate a score.

II. RELATED WORK AND TECHNIQUES

A. Net Promoter Score (NPS)

The Net Promoter Score (NPS) [1][2] is a widely applied technique for measuring customer experience. It is based on asking users in a survey to what extent on a scale of 0 to 10 they would recommend the service provider to friends and family. The users answering with 9 or 10 are considered a promoter, while those answering between 0 and 6 are considered distractors. The score is then the difference of the percentages of promoters minus the percentage of distractors. The core is therefore on a scale of -100% to +100%.

The NPS is a perception score, because it is based on the hypothetical subjective opinion of the user. However, it measures the overall performance of the service provider with respect to user experience. It is not reflecting the perception of a single individual user.

Furthermore, the NPS is entirely based on user surveys. This means that it is a temporary snapshot, only available after a survey is executed. For benchmarking the service provider as a whole, this is perfectly sufficient, but the NPS does not allow understanding momentary changes in perceptions of an individual user. The NPS is undifferentiated and therefore it does not imply any opportunity to learn and understand individual details about a user. Thus, it cannot aid the use cases mentioned in the introduction.

B. Customer Experience Management Index (CEMI)

The customer Experience Management Index (CEMI) is calculating an overall score from various customer experience related metrics [7], collected from all units of a service provider. The calculation method and metrics to be used are standardized in detail, in order to reach good comparability.

Based on measurements within the service provider's BSS/OSS infrastructure, the CEMI can principally be re-calculated at any time, based on latest available data. It does not require execution of user surveys. Nevertheless, the CEMI is still a score assigned to a service provider as a whole. Like the NPS it does not expresses perception of individual users.

C. Mean Opinion Score (MOS)

Mean opinion score is referring to an entire family of scores. Initially it was used in order to capture the perceived quality of voice service sessions. The exact way to assess the quality and generate a score is standardized [3][4]. Users directly classify a presented service quality level. From these user opinions a model can be learned that is able to assess service quality directly from measured performance metrics of the service.

A similar methodology was applied also to other service types. In recent years, there was for example much effort spent on mean opinion score models for video services [5][6].

All MOS are perception scores, usually including a model that calculates a score directly from the objective metrics. The MOS model is however only considering single usage sessions at a time. This is great for assessing the perceived micro-experience of a user as well as the performance of the devices and services involved. It does not allow understanding the overall macro-experience of the user.

MOS as a concept would be applicable to almost all types of services. However, a MOS model has not been investigated for all such services. Also, the maturity and quality of the models vary. If available, a MOS score is used within the SLI model for assessing the quality of micro-experiences.

D. Customer Experience Index (CEI)

The customer experience index is an aggregated experience score of an entire user journey [8]. This means that all kinds of experiences and related scores across all touch-points and channels are considered.

The CEI is defined as a concept. Its description leaves open how each constituent score is derived. Survey based methods are possible as well as predictive scoring algorithms. The SLI can contribute to a CEI by providing a score for service usage experiences.

E. Psychopysics

Psychophysics is the branch of psychology that investigates the dependency of physical experiences and stimuli on human perception, consciousness and sensation [9][10]. It deals with single emotions and thresholds of perception in various environments. Thus it constitutes the base of MOS models and also the SLI.

The laws proposed in psychophysics capture for example the relationship between an external stimulus and individual perception. It also proposes methods to investigate these dependencies. In this respect it is important to note that a simple and universal law that governs the perception of all kinds of stimuli does not exist. This means for the SLI that an individual investigation is needed for every type of service.

III. PSYCHOLOGIAL SCORING

In big data analytics and machine learning the essential step is defining the algorithms, which interpret raw data in order to gain necessary insight. The essential step is the definition of a hypothesis model. The parameters of the model will then be learned based on a set of training data.

This paper introduces a framework of hypothesized models that were created by reasoning about psychological processes of human perception. More specifically, the contributing factors of human subjective perception were identified and transferred into suitable mathematical models.

The overall idea is that the individual experiences of a user, determine his or her satisfaction. The proposed model therefore takes an objectively measured experience of a user, applying a subjective interpretation to it. It is first of all determining how satisfied the user is with the micro experience of single service usage. Then it evaluates how this micro-experience contributes to the overall satisfaction of the user.

This model is trying to re-create the opinion building processes in the mind of a human user, when being exposed to micro-experiences. It captures these processes within a set of scoring algorithms suitable for big-data analytics. Every time the user is experiencing a single event, i.e. micro-experience, his or her macro-experience and with it the related overall opinion is modified. In terms of the scoring model, this means that every time the network is measuring a new service usage experience, the related metrics are evaluated in order to assume how the user's global satisfaction might be impacted by it.

Finding a model that is capturing human perception is a complex task due to the many influencing psychological factors:

- Perception is highly individual. Two users, receiving objectively the same service quality, might still have very different opinions about it.
- Negative experiences are highly significant. Humans tend to take good service for granted in the sense that problems receive higher attention than expected good service.
- Surprising experiences are significant. If users have an experience that is different from what they are used to, it causes a much more conscious reaction. This is true for better than usual quality of experience as it is for worse.
- Memory fades. Older experiences are gradually forgotten and therefore contribute less to the overall opinion than fresh ones.

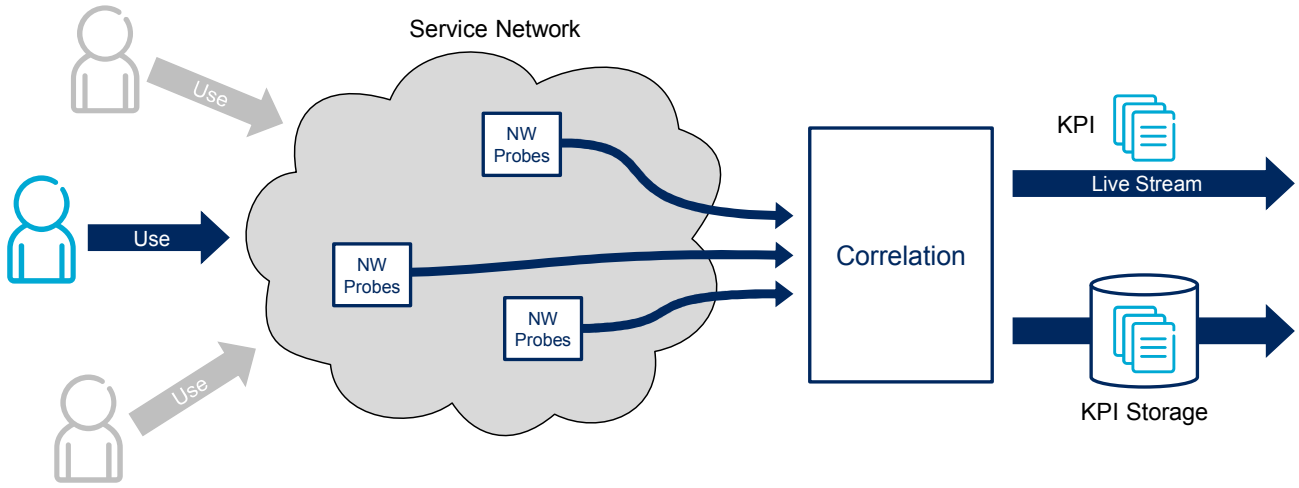

Fig. 1. Collection and Correlation of Network Metrics

- Experiences fade differently. Less significant memories fade quickly while significant experiences contribute to the overall opinion for a prolonged period of time.
- Situation and context matter. The exact circumstances of a service usage can completely change the perception of an experience. Context can for example be the location, date, time of the day, the used device type or the consumed content.

Most of these perception factors express differences in the expectation level of the user. The higher the expectation level of a user related to a particular micro-experience and its context and history, the higher the weight of this experience needs to be in the overall score.

The model described in this chapter interprets user related experience metrics and KPI according to these psychological factors in order to predict the overall satisfaction level of the user. The following sections explain the details of how the model treats individual perception influences, in order to assemble a practical big data analytics model.

A. Correlated User Experience Metrics

The raw data for the SLI scoring is provided by probes in the network. This probe infrastructure identifies service usage sessions and measures performance metrics of these sessions. For example the type of service being used is identified and various metrics that express the service quality are captured. The frame-rate and number of re-buffering events in a video stream or the download latency of websites are typical examples of these measurements.

Considering the millions of users of a service provider and the number of service sessions, the amount of collected data is vast, with high demands on real-time processing capabilities. As early as possible the raw data is reduced by summarizing it into basic key performance indicators, expressing accessibility, retainability and quality of services.

Furthermore, the stream of metrics and KPI is correlated with respect to individual users. The result is a continuous stream of service usage data records. This stream is forwarded for further direct evaluation in real-time or stored in a big data repository for later offline analytics. Fig. 1 shows the basic steps of this measurement process. The Service level index scoring takes this stream of measured data as input.

B. Individualization of the Scoring

When considering human perception, is important to realize that every person is different. A perception scoring algorithm needs to take this into consideration in order to generate meaningful results. It is in particular not enough to find a scoring model for an average user, because the deviation of an individual from the average perception can be substantial. These differences need to be understood and managed in the scoring.

Individualization does not necessarily lead to different calculation algorithms as such. Every user has for example a long-term and a short-term memory of experiences, but it is a very specific property of a user if a particular experience would be remembered for a long time or quickly forgotten. This means that the parameters for model variables, guiding the scoring, would vary for individual users, while the algorithm as such is the same for all users.

Individualization means that a personal set of model parameters is assigned to every user. These parameters tell the scoring algorithm how to interpret the raw data for this particular user. This practically means it is necessary to find and manage as many sets of parameters as there are users.

Finding model parameters requires investigating in detail how a user would react to certain service quality levels. This is usually done by means of surveys asking for a user's opinion, combined with the measurement of the objective service quality metrics presented to the user. This is an extensive investigation needing collaboration of the user. It is practically impossible to execute this for each of the millions of users of a typical service provider.

While there are major deviations from an average user's perception it still possible to find groups of users, who are quite

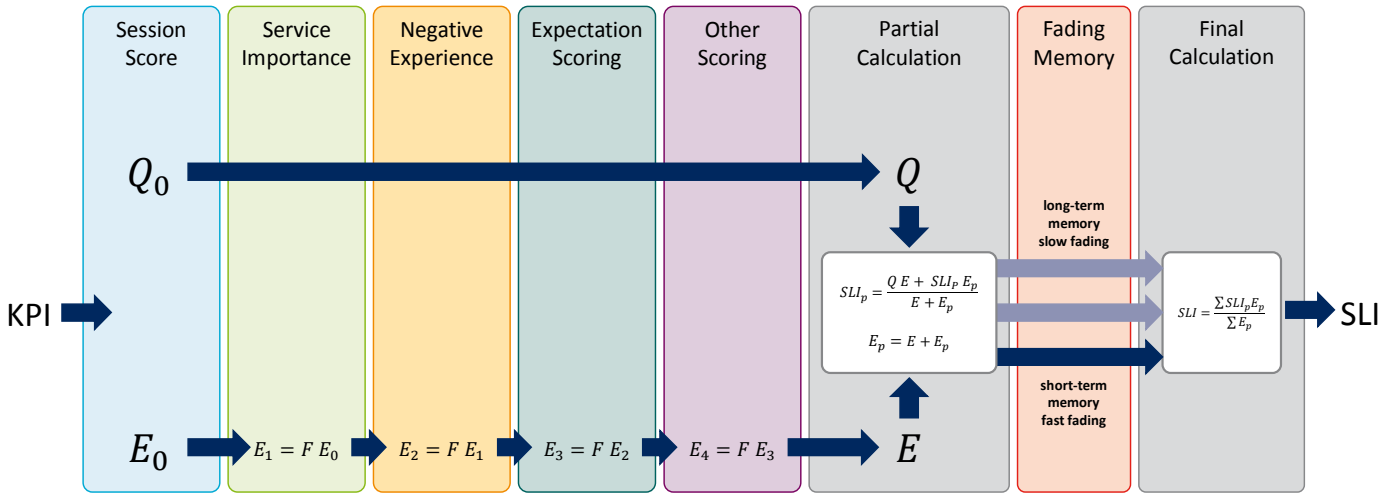

Fig. 2. The Scoring Pipeline of the Service Level Index

similar in the way they perceive services. For example users with a high consumption of streaming videos can be considered to be in general more sensitive to video quality. It should therefore be possible to find a set of model parameters expressing the typical perception of users with high video consumption. This can be achieved by doing the survey based investigation only with users from the high video usage group.

Group specific investigation can be done for a couple of different user groups. Various criteria can be used in order to define these groups, as long as the distinction implies a significant difference in service perception. Service usage levels are a good criterion. They distinguish users with different needs and expectations, implying differences in perception and opinion. Other feasible criteria for user grouping are types of subscriptions or demographical factors like gender, age and location. Business premium users and private pre-paid users typically have different needs, leading to different perceptions of the delivered service quality.

The user groups for perception scoring are not necessarily the same as user segments for marketing and portfolio planning purposes. However, in practice, both are often very similar to the extent that existing user segmentation could be reused also for perception scoring.

Every user is in general assigned to several user groups. For example a teenager, with a pre-paid subscription and high online gaming usage is assigned to three respective user groups. This particular user is then considered to have a mixed perception typical for these three groups. The individual set of model parameters for this user is then calculated from the reference parameters sets assigned to these groups. The default method for combing group specific reference parameter sets is calculating average values over all contributing sets. This means a user gets a personal set of scoring parameters according to his or her individual user group memberships. Only two users, who are in exactly the same user groups, will then have the same perception scoring setup.

Employing user grouping as base for individualizing perception scoring is practical and by that feasible, because it requires investigating a model parameter set per user group, rather than per individual user. It is a tradeoff between the most accurate individual models and the use of a single model parameter set of a globally average user.

Individualization by this method is also possible for users, who have never been part of a user survey. If users are members of the same user group this is considered to imply similar perception. Survey based parameterizations are extrapolated to all members of the group.

This group based individualization method might still be inaccurate for some users, but experiments have shown that it allows statistically much improved results compared to using a global model for all users. It is has proven in practice to be a good tradeoff between impractical full individualization and low accuracy of global model parameters.

C. *Scoring Pipeline*

The high number of influencing factors on perception suggests a model with a high number of variables. This means that it is not feasible to simply machine learn a scoring model. There is not enough significance in the raw data for such an approach. Simplification of the model by removing variables is also not a good idea, because it means neglecting factors that psychological research has found to be relevant influences.

The solution to this dilemma is separating concerns. Each influencing perception factor is treated within a separate model with dedicated mathematical description of its influence and its own set of configuration parameters. All sub models are then combined into an overall frame for determining the combined score.

A pipeline structure is chosen as shown in Fig. 2, in order to allow each sub-model to contribute its results to the overall score. The scoring pipeline handles one experience event at a time. Its associated metrics are the input and a new score is generated. Following this, the next event and metric is fed into the scoring leading to another update of the score.

Along the pipeline two values are collectively determined by all sub-models: The quality Q is the subjective quality score

of a micro-experience. The emphasis E expresses the significance of this event in the context of the macro-experience. In this respect the overall score is a weighted average over the perceived quality of all micro-experiences, where the emphasis constitutes weight.

D. Basic Quality Score

The first step in the scoring pipeline is to evaluate the perceived quality and find the value of Q. The KPI or metric of the experience event is taken as input and a scoring function then determines the associated quality score.

MOS models are able to do exactly this determination of a service session quality from raw metrics. If a MOS is available, it can therefore be directly used in this step of the SLI model.

With an emphasis initialized to the value 1 the lowest significance is assumed by default at the start of the pipeline

E. Service Type Importance

Every user has favorite services and services that are highly important to them. For example, a user, who is often attending business meetings over the phone, will pay a lot of attention to voice quality, accessibility and retainability. An online gamer on the other hand might be very sensitive to data latency while voice is not that important for a good experience.

The sub model for service importance considers the general importance of a service to the user. It does this through applying a factor to the value of the emphasis for those service types that are assumed to be important to the user.

F. Negative Experience

It is a known from psychological research that a negative experience has a much bigger impact on human perception than a positive one [12]. A sub-model of the SLI scoring is entirely dedicated to handle this effect. The quality score is taken as a measure of how negative the experience is. The scoring logic then determines an emphasis factor that is higher for lower Q values.

G. Surprizing Experiences

Users are accustomed to the level of quality they usually receive in their service usage. Even if the quality is not good, it is the one they are familiar with. If the services show a sudden change in quality, this is noticed particularly and has to be treated with higher emphasis.

In order to express this mathematically a measure of typical level of quality is needed. The current SLI score of the user is expressing exactly this. The measure of surprise in service quality is then the difference between the Q value assigned in the current scoring and the long term SLI score. This is then transformed into another emphasis factor.

H. Other scoring sub-models

The pipeline structure of the overall scoring allows a high degree of separation of individual psychological factors. This architecture also allows for very easy extension of the model. There are surely more psychological factors that play a role in user perception than included thus far. Once another factor is sufficiently understood, it can be expressed in yet another dedicated sub-model and added to the pipeline.

I. Fading Memory

Fading of memory means the continuous reduction of weight of older experience in the overall score. The most obvious implementation would be through a rolling window where all data older than the timeframe of the window is not considered any more. This method would treat all data equally and all experiences are forgotten at the same rate. This is not how the human memory works. It is in fact highly selective with respect what and how fast an experience is forgotten [11]. A major rule is that a memory is kept longer, the more significant the related experience was.

Keeping fully track of each experience event with an individual fading factor would be flexible enough in order to manage individual fading. However, this kind of solution would mean a high degree of data management effort for each user with high demands on memory consumption and processing resources. For real-time scoring especially, this would not be an ideal scenario.

The proposed method introduces several memory lanes representing short-term and long-term memory and a number of different mid-term memories. The idea is to manage partial SLI scores individually in each of the memory lanes and have a different decay of significance.

This process can be understood considering that each SLI score is a weighted average of quality scores, with emphasis used as weight. Assigned to it is the sum of all emphasis used in the past. When scoring a new experience, the SLI score is updated by calculating the weighted average of the new quality value, with its new emphasis and the old SLI score with its associated emphasis sum.

$$SLI_n = \frac{Q_n E_n + SLI_{n-1} E_{sum}}{E_n + E_{sum}}$$

The emphasis sum expresses the weight of old experiences compared to the weight of new experiences. Reducing the emphasis sum therefore means a decay of the old experiences. It effectively means forgetting. A controlled decay of the emphasis sum value is the mechanism to create a fading memory.

Several memory lanes are needed in order to distinguish experiences that shall be forgotten quickly, from those that shall contribute to the overall score for a longer period of time. Every memory lane uses the formula above to generate a score and it decays the emphasis sum over time. The difference between the memory lanes is solely the decay rate.

When a new experience is scored it needs to be decided into which memory lane to place it. Psychology shows that an significant experience is remembered longer. The emphasis is a measure of this significance. This means, the higher the emphasis value, the more long-term the chosen memory has to be. A set of thresholds controls this decision.

The result is a partial SLI score and emphasis sum per memory lane where the emphasis sums changes over time even if there is no new scoring. The final SLI score can be calculated at any point in time as weighted average over the partial SLI scores in the memory lanes.

This way of implementing memory fading means that the scoring will never need to return to historic scores or even raw data. This makes the model suitable for efficient big data analytics processing and even real time scoring. Nevertheless, it allows for fading memory of significant and insignificant experiences at different rates.

IV. Properties and Handling of the Model

The modular nature of the scoring model means that new sub-models can easily fit into the scoring. The scoring module interface is simple. A module needs to accept a Q and E value pair as input and forwards a new and potentially modified Q and E pair to the next scoring module. In principle the quality and the emphasis score can both be modified along the scoring pipeline of the model. In current practice the quality score is left as it is found by the basic session scoring step, while the other currently existing sub models contribute to the emphasis solely.

The basic session scoring is similar to MOS models and uses a MOS model directly, if a mature one is available for a service type. The rest of the scoring modules are concerned with the question of how significant this currently scored micro-experience is, in the context of all past experiences of the user and thus how much impact the related single quality score would have on the user's global satisfaction.

The SLI model is able to operate in real-time stream processing mode, where all metrics are fed into the pipeline directly from the correlated stream of probe data. It is also possible to implement it for batch processing. In both cases it would produce the same score value.

It is often not the absolute value of the SLI that is interesting, but the detection of change in the score. For example the service provider might use the SLI in order to identify users, who recently became dissatisfied. This is shown by a recent and significant drop in the associated SLI value. These users can therefore qualify for remedial actions, such as compensatory promotions.

Another possibility is to further investigate the cause of the score drop. The experience history would then be utilized for deeper investigation, in order to identify any potentially systematic problems with the service environment. This would allow an early identification of issues, while they still affect only a few users. Promising targets for infrastructure investments with a clear impact on user experience can also be identified.

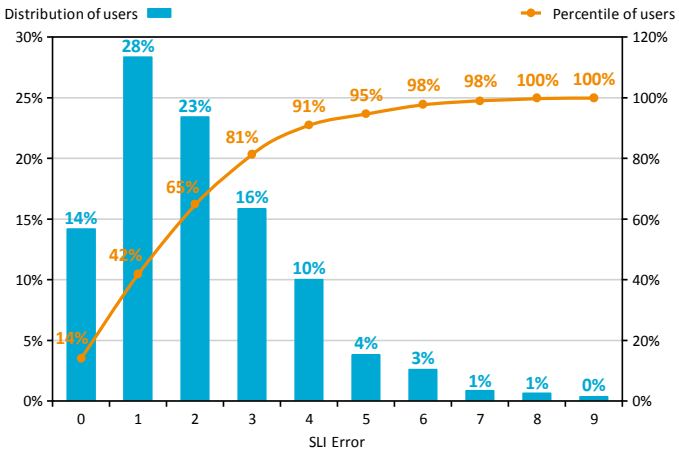

Fig. 3. Verification of SLI prediction against NPS survey answers

V. Root Cause Analysis Support

Having a perception score prediction is a huge step forward for individualized management of user experience and individual reactive and proactive actions by the service provider. However, the score alone only shows the momentary satisfaction of the user, but it does not help understanding the reason for the assumed level of satisfaction. Root cause analysis support was introduced in order to gain further insights. It answers the question, why user's satisfaction is what it is, or why it has recently increased or decreased. Major changes in the score are often caused by a small number of highly significant experiences. Especially errors that temporarily make a service unusable are major events with high impact on the score.

Information of the major scoring events is preserved, in order to support any review of the user's score. A list of the top scoring events of the day is generated at the partial SLI calculation step in the scoring pipeline. At that step the individual scoring of single experiences is finished and because of that a final emphasis score is determined.

It is possible to choose other time intervals than single days for preserving the data of the top contributing experiences. It is also possible to combine several daily lists in order to assess longer time-frames. Filtering these lists while still using the emphasis as sorting criterion can also lead to further insights. It is for example possible to see if particular service types have contributed disproportionally to the long term macro-experience.

VI. Verification of the Prediction Quality

The service level index is a predictive score. For verification of the prediction accuracy it needs to be compared with a reference value. In this case the reference is the user satisfaction, as directly stated by the user. This means the verification is based on surveys.

The service level index scoring was applied to and verified in a live network, as part of an extensive proof of concept project. Verification is done by calculating the SLI score for a set of users based on their service usage measurements. This calculation was continuously executed for two weeks. At the end of the observation and calculation period the same users were then asked in a survey about their satisfaction with the operator.

In the end there are two values available for each participating user. The direct answer to a satisfaction survey constitutes the reference. The prediction error for each individual user is the difference between this user's survey answer and the SLI score. The overall quality of the prediction model can be read from the distribution of these differences.

The result of the first verification project based on real network data is shown in Fig 3. The predictive model shows a

significant accuracy in the sense that for the majority of users the error is small and in the magnitude of one or two scoring points. Huge errors, where the prediction score is totally wrong exist, but they are comparably rare.

Deviations in the predictive score cannot be completely eliminated due to influencing perception factors that are currently beyond measurement. Overall the perception scoring model appears to be good at predicting at least the correct tendency for the majority of users.

This is an encouraging result especially because a simplified model was used in this early trial with rudimentary calibration. After these experiences both the scoring models and the calibration process were considerably extended and refined. At the time of writing this paper further proof of concept projects are executed and first commercial applications are installed.

In commercial operation it is not necessary to do extra user studies for SLI verification. Those studies done for determining a net promoter score can be used also for verifying the SLI prediction accuracy. This means a continuous prediction quality assurance can be reached with low additional costs.

VII. CALIBRATION OF THE MODEL

The calibration of the scoring is highly important for good prediction results. It is about finding the exact model parameter values. For calibration user studies are necessary because they make up the reference point.

As the individualization of the model is based on user grouping, a separate calibration is done per pre-defined group of users. This means that a sufficient number of users need to participate in the study in order to have enough reference points per user group.

Due to the complexity in the domain of human perception and the number of variables that are still in the model, several calibration techniques are combined. A basic concept is separation of concerns. This means that sub-sets of model parameters are calibrated separately from others. The sub models in the scoring pipeline provide already a good modularization for this purpose.

The first calibration technique is used for example with the basic quality scoring sub-model. Participating users are asked to provide their opinion about single service sessions. From these answers and with the help of a regression analysis a sub model is calibrated that translates a metric into a quality score. This is a process similar to MOS model investigations.

Using per sub-model calibration methods can find good values for many free parameters of the overall model, but not for all such parameters. For example, the thresholds for memory fading are difficult to separate from the rest of the scoring model. For remaining parameters similar to these, an iterative overall calibration process is executed. It calculates SLI scores for all participating users with a set of parameter values and then verifies the overall SLI accuracy as described in Chapter VI. Then the parameters are modified and a new SLI is calculated. This is the basic idea of a search process that is repeated many times until the configuration with the best prediction result is found.

The description of this calibration process as provided here is a simplified version. It is basically a search for the configuration with global minimum of prediction errors. In this respect, the first step of a separate calibration per sub-model was done in order to considerably reduce the dimensions of the search space, while approaching the prediction error minimum. This combination makes the calibration process feasible in a practical way, despite the complexity of the domain.

Also here, the considerations from psychological research are used in order to eliminate model parameter value ranges that do not make sense in the context of human perception. This way the number of free variables and configurations to be tested is decisively reduced.

VIII. PERCEIVED QUALITY OF NETWORK ASSETS

The psychological scoring model of the SLI is mainly created for assessing a user's perception and the user's overall satisfaction. For optimizing infrastructure investments with the goal to improve user experience it would also be beneficial to understand how a particular network assets contributes to the experience. These assets can be any network node like application servers or radio cells.

A perception score of a network asset expresses how all users, who were served by that asset, have perceived the service quality provided with that asset being involvement. This would mean the operation and performance of an asset can be monitored by means of the subjective quality it provides to users, rather than only by its objective technical metrics.

This idea was first implemented for radio cells. The respective score is called cell level index (CLI). It is generated by taking all experience metrics and KPI from service sessions where the scored radio cell was involved. This stream of experience data is then fed into the SLI scoring. The ID of the radio cell, rather than the user, is the main key for correlation and filtering.

This means that the scored data stream represents a mix of all users served by that particular radio cell. The score for the radio cell becomes a perception score of a virtual user, which is constructed from the many individual experiences of all the users served by the cell. Each of the individual users has their own personal perception expressed by their own SLI model parameter set. This is taken into consideration when calculating the CLI by loading the parameter set into the scoring model that belongs to the user, who has actually experienced the currently scored input metric. The scoring parameterization might therefore change dynamically with every new metric fed into the scoring pipeline.

This method can not only be used for scoring a perceived quality of radio cells, but in general it can be used for any network asset. For example the performance of media streaming servers can be evaluated this way. The only prerequisite would be the availability of identification data of the network asset in the experience data set.

IX. Summary and Outlook

Big data analytics usually means to find the right interpretation of data. A huge arsenal of empirical methods is available for finding and calibrating respective interpretation algorithms and models. Machine learning and various statistical methods are used. This usually also means that, based on a set of training data, significant insights about a given problem need to be derived from that data. Mathematically this means that the parameters of a hypothesized model are optimized for best fit to the training data set.

The art in machine learning is to find a hypothesis that is complex enough to accurately capture all aspects of the wanted prediction, in as many situations as possible, while still finding enough significance in the training data for actually learning the related model parameters. In the given domain of predicting satisfaction from network measurements, the amount of raw data from the network is vast. As the capability for real-time processing of this data is required, the prediction model needs to be suitable for scalable stream processing. The proposed model scales well with increasing numbers of processing nodes.

While there are huge amounts of raw data available, generating training sets requires complementing this data with correlated user survey results. This is an expensive step and thus data sets suitable for training the model are sparsely available. This translates directly into a limitation on the number of independent parameters the model can have.

Psychology was used in order to include additional insights, aiding the reasoning about a good model hypothesis that is as complex as necessary, while staying simple enough to be learned from practically available training sets. However, when reasoning about human perception and human satisfaction, the amount of potential influencing factors is infinite. The available training data does not provide enough significant insights to support a decision of which of these factors is significant. It turned out that a reliable and generally applicable model for perception scoring could not be found by just formally learning from sets of historical training data.

Nevertheless psychology and its insights about human perception became the key for finding a practical model. It first of all allowed us to reason about which influencing factors on perception are more significant than others. Based on this, the model hypothesis could be built focusing the most significant variables representing the important factors.

Furthermore, psychology driven reasoning also helped narrowing down value ranges of variables and pre-selecting certain functional dependencies between variables. In practice, many mathematically possible options to describe the influence of a variable could be dropped, because they would express human perception processes that do not comply with experiences from psychology. In this respect it should be noted, that each variable in the model and its values directly expresses and quantify a psychological factor. This allows to directly reason about the results of the learning process.

The result of the effort is a model for big data analytics that provides a good interpretation of the data while considering the most significant influencing factors. Together with the model an elaborate process and methodology for learning and calibrating the model parameters were developed. This calibration allows customization of the model to any given user base for optimized accuracy.

However, still not all influencing factors can be captured by a predictive model based on network measured data. Some influences and its causes do not show in this data source. For example the influence from the non-virtual social environment, also known as “real life”, is hard to capture and often completely out of reach for data analytics. An important future direction is therefore the extension of the scoring to further sources of experience other than network provided services. Examples are experiences of the user, when visiting a point of sales or calling customer care. Ultimately all touch-points and all channels between the user and the service provider will be covered. Continued research is done in order to understand these factors and capture them into an extended scoring model. This will lead to an even higher compliance with psychological perception processes, resulting in higher accuracy of the score. In order to allow these future extensions, the model was built to be modular and extensible.

So far, the existing perception scoring model has proven to provide an essential input to many mission critical processes within a service provider’s organization. Marketing and business planning can gain from detailed and individual understanding of their user base. Network operation can better understand how network assets impact user experience and optimize investments respectively.

References

[1] F. Reichheld, "One Number You Need to Grow" Harvard Business Review, December 2003

[2] F. Reichheld, R. Markey, “The Ultimate Question 2.0: How Net Promoter Companies Thrive in a Customer-Driven World.”, Harvard Business Review Press. p. 52, 2011, ISBN 978-1-4221-7335-0.

[3] ITU-T Recommendation P.800, “Methods for subjective determination of transmission quality”

[4] L.A.R. Yamamoto, J.G. Beerends, “Impact of Network Performance Parameters on the End-to-End Perceived Speech Quality”, In Proceedings of EXPERT ATM Traffic Symposium, 1997

[5] S. Winkler, "On the properties of subjective ratings in video quality experiments", Proc. Quality of Multimedia Experience, 2009.

[6] ITU-T Recommendation P.910, “Subjective video quality assessment methods for multimedia applications”, 2008.

[7] TM Forum Best Practices RN341, “Customer Experience Management Index (CEMI)”, Frameworkx version 12.5

[8] TM Forum Frameworx Best Practice GB962, “Customer Experience Management, Introduction and Fundamentals”, Release 14.5.1, March 2015

[9] G. Gescheider, “Psychophysics: the fundamentals“ . 3rd edition, 1997. Lawrence Erlbaum Associates. p. ix. ISBN 0-8058-2281-X.

[10] K.R. Boff, L. Kaufman, J.P. Thomas (Eds), “Handbook of perception and human performance: Vol. I. Sensory processes and perception”, John Wiley, New York.

[11] Jr. Murdock, B. Bennet, "The serial position effect of free recall." Journal of experimental psychology 64.5 (1962): 482.

[12] Ito, Tiffany A., et al. "Negative information weighs more heavily on the brain: the negativity bias in evaluative categorizations." Journal of personality and social psychology 75.4 (1998): 887