Data-driven Insights through Visual Analytics

Empowering Innovative Data Solutions

kpmg.ch

Data Science and Machine Learning approaches have been in rapid development, quickly becoming a key success factor for companies. At the same time, high complexity and multi-variate features pose significant challenges in the analysis of common datasets. To achieve business success, the application of state-of-the-art approaches is crucial for gaining competitive advantages. A key success factor is to turn data insights into knowledge in order to fuel business actions. Applying Visual Analytics principles in this process creates a powerful symbiosis of Data Science approaches and the unmatched ability of humans to pre-attentively detect patterns, transferring insight to knowledge and knowledge to actions.

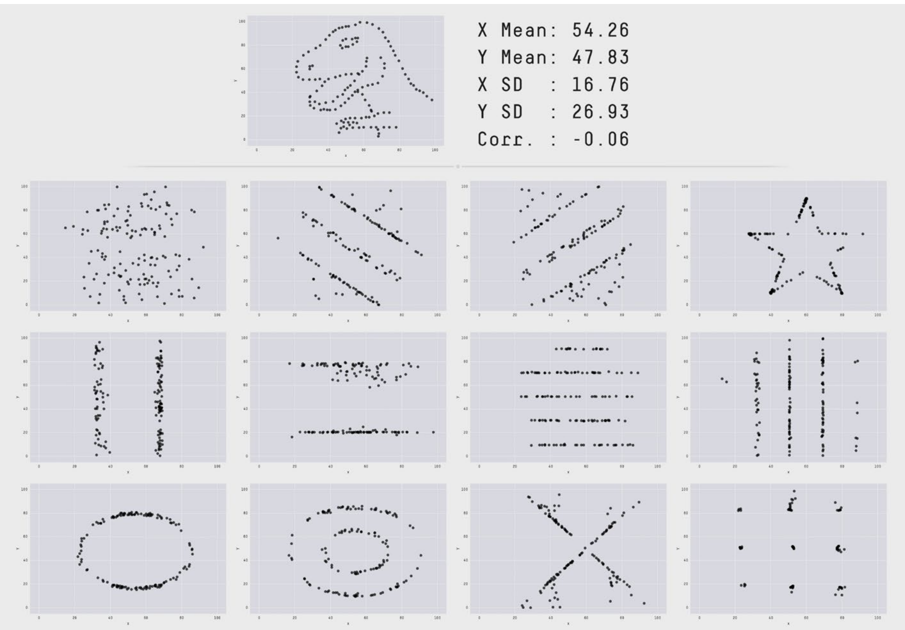
The availability of digital processable data, such as financial data, news, textual communication or customer data, is increasing. Data and Analytics are nowadays pervasive in all aspects across different business areas and become even more prominent and essential through the ongoing digital transformation across industries.

Applying purely statistics does not provide sufficient insight into data

Data is mostly highly complex and multi-variate. Considering customer data as an example with extensive demographic information, financial transactions or insurance data, these examples show the whole complexity of a high-dimensional data space. To get a first understanding of data, prescriptive statistics usually offer good starting points, for example by calculating means and standard deviations of customer age, or observing the distribution of nominal categories as, for instance, gender. But statistical analysis alone does not transport the full picture on a dataset and thus does not suffice to gain extensive knowledge about a dataset.

To exemplify this coherence, in 1973, the English statistician Francis Anscombe created four datasets with nearly identical descriptive statistical properties, but strongly distinctive value distributions in each set of the 'Anscombe's quartet'[1]. In 2017, Matejka and Fitzmaurice[2] presented a simulation prototype that extended the idea of the 'Anscombe's quartet', generating datasets with the same statistical features over many iterations, as shown in Figure 1. The authors start each simulation run with the Datasaurus dataset, created by Alberto Cairo[3], and show each iteration step in an animated fashion to understand how the final dataset evolves – all iterations having the same statistical measures.

The clearly completely different value distributions would go undetected if only explored by statistics – only looking at the distributions reveals differences and structures that need to be considered for decision making.



> "The purpose of computing is insight, not numbers."
>
> **(Richard Hamming, 1962)[4]**

Fig. 1. Same statistics – different visual graphs[2]: Even though the datasets differ in their visual appearance, they all share the same statistical measures, as mean, standard deviation and correlation.

Consequently, as Hamming stated in a famous quote "The purpose of computing is insight, not numbers"[4], he expressed the necessity to enrich scientific research approaches using visual techniques to derive new and more complete insights in large data collections.

Visualization as knowledge compression

Visualization enables humans to get insight into data, which can be viewed as "knowledge compression". A huge amount of information and knowledge can be represented in a relatively small space, thus enabling humans to easily interpret the information and to increase the analysis speed. While simple visualizations provide good and knowledgeable insights, many remain rather static and do not react to changed input data.

For instance, in Financial Services, stock markets are changing constantly, and new input data is streamed into data warehousing or other storage systems. To explore the data and react to changes, static visualizations are not sufficient.

> "Visual Analytics combines automated analysis techniques with interactive visualizations for an effective understanding, reasoning and decision making on the basis of very large and complex datasets."
> **(Keim et al. 2010)** [5]

Visual Analytics – powerful symbiosis of Data Science approaches and human perception

To gain knowledge from insights, a powerful concept is the combination of Data Science approaches with the unique human visual perception and business experience, named "Visual Analytics".

Schematized in Figure 2, the Visual Analytics methodology propagates the integration of automatic and visual-interactive data analysis, thus combining the unmatched cognitive abilities of a human with the precision of computer-based Data Mining methods. The term **Visual Analytics**, or shortly **VA**, is defined by Keim et al. as follows: "Visual Analytics combines automated analysis techniques with interactive visualizations for an effective understanding, reasoning and decision making on the basis of very large and complex datasets"[5].

Through interaction, data can be explored visually, and patterns or anomalies can be pointed out, leading to new insights and knowledge.

Knowledge generation using Visual Analytics to gain insights into data, and to recommend action

The **Knowledge Generation Model** for Visual Analytics by Sacha et al.[6] describes a model for the knowledge generation process in Visual Analytics systems. The authors present a model that divides a system into a computer-aided and a human part.

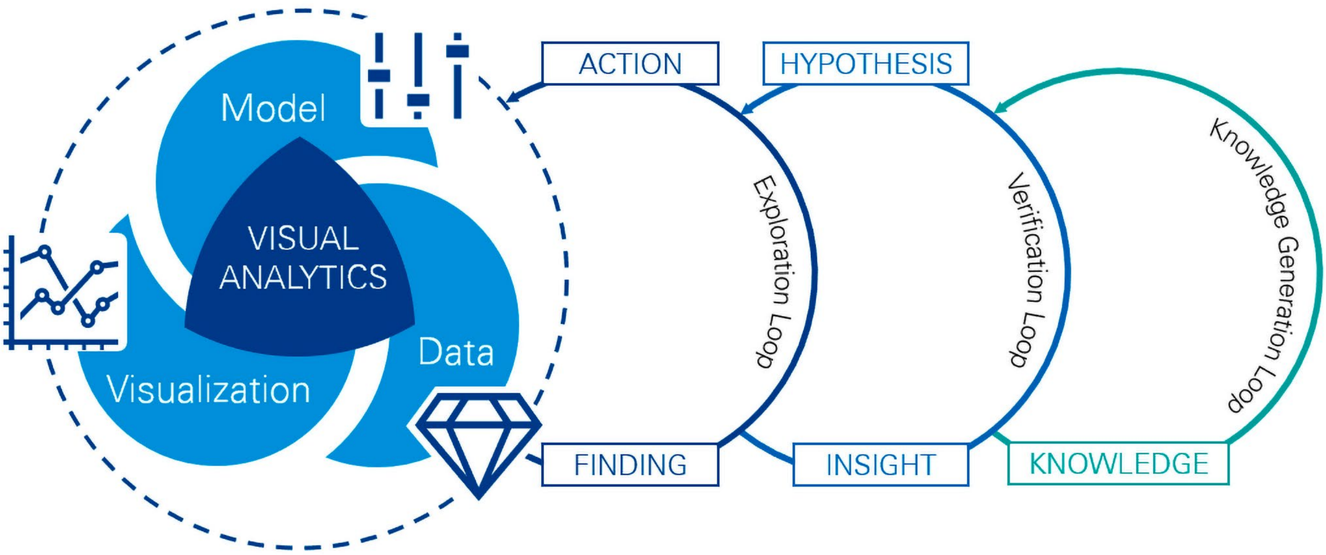
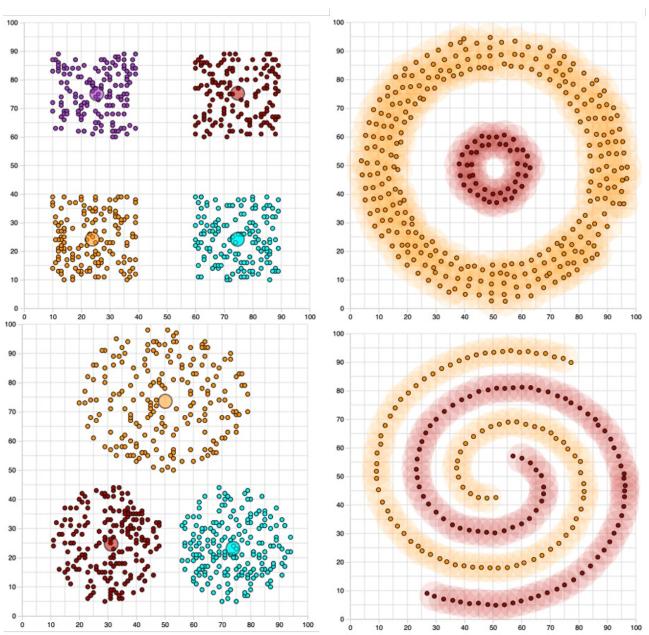


Fig. 2. Visual Analytics combines the integration of automatic and visual-interactive data analysis approaches with the unmatched cognitive abilities of humans to derive knowledge and gain insights into data.

The computer-aided part represents a VA system, as shown in Figure 2, and consists of three parts: Data, Model and Visualization. In our use case, a VA system analyzes data with the usage of Data Science techniques, which refer to the model. These approaches can reach from classical Data Mining techniques, such as clustering or feature analysis, to Machine Learning models, as Random Forests or Neural Networks, up to Deep Neural Networks. Most of the models create statistical results, that are expressed by various statistical measures, e.g. accuracy or errors. With the support of visualization, the modeled results can then be explored interactively by humans with their unique ability to perceive patterns and small changes at a glance.

The approach to generate knowledge is moving from the classical human-in-the-loop to the human-is-the-loop strategy [7], who actively controls this process. In this concept, see Figure 2, there are three abstraction levels that a user passes through during knowledge generation.

The basic **Exploration Loop (EL)** maps simple interactions with the system, such as parameterization of the selected Data Science model and animation of the analyzed results. One level deeper is the **Verification Loop (VL)**, in which the user can test within the Exploration Loop hypothesis or generate new ones. The totality of the hypotheses formed and verified in the previous loops now result in the knowledge the user has gained through the **Knowledge Generation Loop (KL)**.



Centroid-based clustering: K-means

Density-based clustering: DBSCAN

Visual Analytics for fraud detection in Financial Services

In the case of financial transactions, new dangers are brought by the digitalization of the monetary sector, in which banks must increasingly address issues concerning data security, data management, data governance and data protection.

Methods for fraud detection that have proven successful at times need to be adapted increasingly fast to changing threat vectors and heterogeneous attacks. This scenario provides a perfect use case to illustrate the symbiosis of computer-aided, but human-controlled, data analysis to detect patterns and anomalies in transactional data. Data Analytics can reveal such patterns and anomalies in large data collections.

A crucial aspect in the case of fraud detection is that these insights must be available in real-time so that immediate action can be taken in the event of fraud. An intelligent fraud detection system would alert when suspicious transactional movement occurs which contradicts normed patterns.

But in anomaly detection, it is important to identify which transactions correspond to normal behavioral activity and which transactions fall outside this pattern in the first place. The application of anomaly detection algorithms, which results in single scores, is not sufficient to reveal the complexity behind complex and sophisticated fraud attempts, which might go unrecognized by algorithms that were never exposed to new fraud strategies.

Clustering reveals anomalies

Clustering approaches can be applied to do so by not only revealing outliers and anomalies but by giving visual insights into the data structures and patterns to be found within the data sets. By visual exploration, and the application of center- or density-based clustering methods, analysts can dive deeper into potential fraudulent cases.

Two prominent clustering examples for center- and density-based clustering approaches are displayed in Figure 3. On the left-hand side, k-means [9] clustering detects four clusters in the upper example and three in the lower scenario.

As can be seen, the number of clusters needs to be known beforehand, otherwise the algorithm merges distinct groups or splits cohesive data points apart. For comparison, on the right-hand side, the DBSCAN algorithm [10] is shown as a representative for density-based clustering.

Fig. 3. Exemplary illustration of various datasets and applied clustering methods, to demonstrate the applicability to different data structures. Clustering examples were generated using EduClust [8] – a visual education platform teaching clustering algorithms.

Neighboring data points are assigned to the same cluster, forming different shapes that represent the underlying data structure. More complex data structures can be clustered with more sophisticated and advanced Data Mining approaches for clustering, such as t-SNE[11], UMAP[12] or parameter-free hierarchical clustering techniques as FINCH, which was recently introduced by Sarfaz et al.[13].

This exemplary scenario shows the importance of the application of visualizations to explore the results of Data Science approaches. It is crucial in fraud detection to not only detect outliers but also to choose the right clustering approach that matches the underlying data and its characteristics. Choosing the right approach depends on data distribution, use case and user experience and should never be taken lightly. Visual Analytics approaches help in choosing suitable approaches through the combination of interactive visual and statistical verification.

Another important factor is not only the choice of the best-suited classifying approach, but also being able to trust the results by having robust and validated techniques that commonly identify threads and lower uncertainty.

The wrong parameterization of the chosen algorithm can also lead to falsely classified threats, resulting in incorrect assessments of threat levels which can lead to missed assaults. In this example, an interactive Visual Analytics prototype that combines Data Mining approaches and the user-centered parameterization of the models provide comprehensive features for a fraud detection system as robust as possible against these effects. Threats can be detected automatically and explored visually by the user to gain insights into the attack, as well as to inspect the fraudulent transactions to finally update and broaden the knowledge base.

Design Thinking to develop Visual Analytics systems

The design and development of a Visual Analytics system is a user-centered process and requires not only business and technical knowledge but also an understanding of the clients and business needs.

Therefore, a Design Thinking approach, as shown in Figure 4, is apposite to understand the business needs to translate them into a Visual Analytics prototype ideation with well-suited Data Science methods and approaches to gain insights and create value out of data in a human-centered manner.

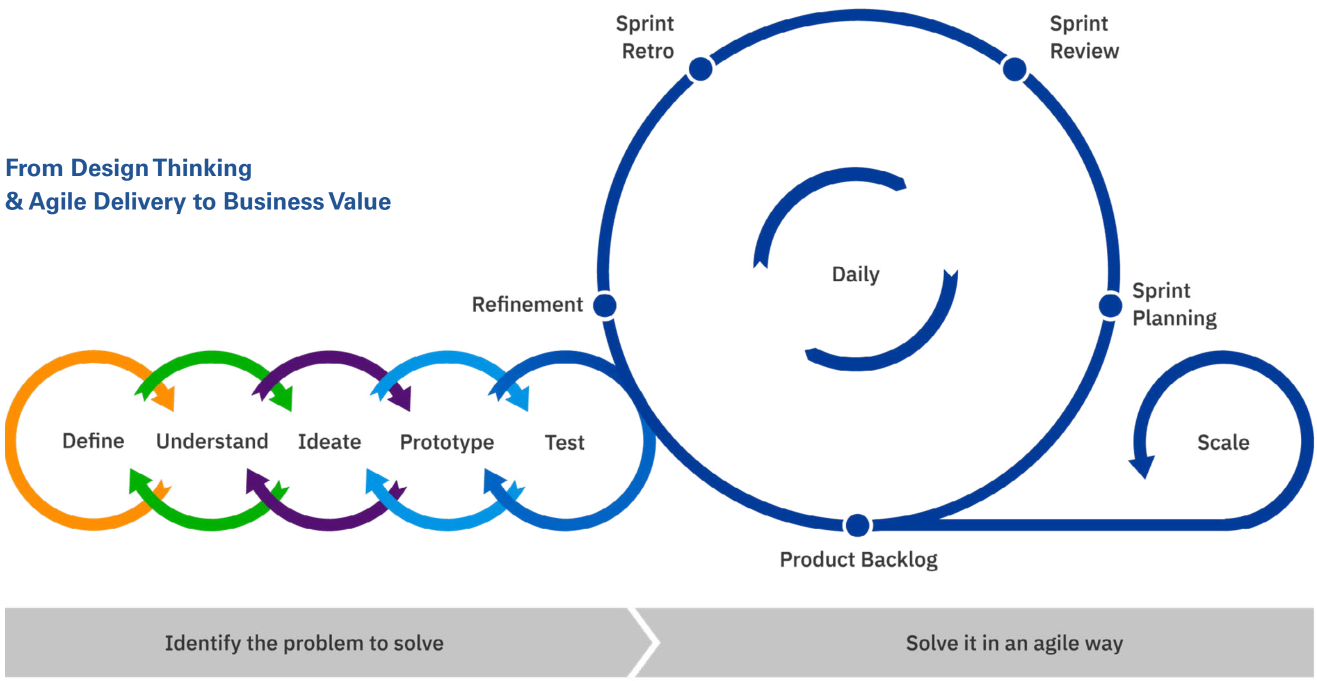


Fig. 4. The convergent journey from a user-centric approach into an agile base delivery approach

Conclusion

Data-driven innovation is a key differentiator in a competitive market environment and can set a business ahead. A reliable partner is needed in order to create tailored solutions and combine sophisticated Data Analytics and state-of-the-art Data Science approaches and technologies with human-centered interactive ideated solutions for creating impact and value out of data. Please do not hesitate to contact us should you wish to receive more details on how KPMG can help your business to develop Visual Analytics prototypes to fully uncover the potential of your data.

Bibliography:

[1] Anscombe, Francis J. "Graphs in Statistical Analysis" The American Statistician 27.1 (1973): 17-21.

[2] Matejka, Justin, and George Fitzmaurice. "Same Stats, Different Graphs: Generating Datasets with Varied Appearance and Identical Statistics through Simulated Annealing" Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems 2017.

[3] Cairo, Alberto. "Download the Datasaurus: Never Trust Summary Statistics Alone; Always Visualize Your Data." http://www.thefunctionalart.com/, Alberto Cairo, 29 Aug. 2016, www.thefunctionalart.com/2016/08/download-datasaurus-never-trust-summary.html.

[4] Hamming, Richard. "The purpose of computing is insight, not numbers." Numerical Methods for Scientists and Engineers (1962): 249-264.

[5] Keim, Daniel, et al. "Mastering The Information Age – Solving Problems with Visual Analytics" (2010). Conference on Computer Vision and Pattern Recognition. 2019.

[6] Sacha, Dominik, et al. "Knowledge Generation Model for Visual Analytics" IEEE Transactions on Visualization and Computer Graphics 20.12 (2014): 1604-1613.

[7] Endert, Alex, et al. "The Human is the Loop: New Directions for Visual Analytics." Journal of Intelligent Information Systems 43.3 (2014): 411-435.

[8] Fuchs, Johannes, et al. "EduClust: A Visualization Application for Teaching Clustering Algorithms." Eurographics 2019-40th Annual Conference of the European Association for Computer Graphics. 2019.

[9] Hartigan, John A., and Manchek A. Wong. "Algorithm AS 136: A K-Means Clustering Algorithm." Journal of the Royal Statistical Society, Series C (Applied Statistics) 28.1 (1979): 100-108.

[10] Ester, Martin, et al. "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise." KDD. Vol. 96. No. 34. 1996.

[11] Van der Maaten, Laurens, and Geoffrey Hinton. "Visualizing Data using t-SNE." Journal of Machine Learning Research 9.11 (2008).

[12] McInnes, Leland, John Healy, and James Melville. "Umap: Uniform Manifold Approximation and Projection for Dimension Reduction." arXiv preprint arXiv:1802.03426 (2018).

[13] Sarfraz, Saquib, Vivek Sharma, and Rainer Stiefelhagen. "Efficient Parameter-free Clustering using First Neighbor Relations." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019.

Contacts

KPMG AG
Badenerstrasse 172
P.O. Box
CH-8036 Zurich

kpmg.ch

Thierry Kellerhals
Director
Lead Digital Experience
Financial Services

+41 79 281 22 50
tkellerhals@kpmg.com

Isabel Piljek
Consultant
Data Scientist
Financial Services

+41 76 261 52 39
ipiljek@kpmg.com