Using Data and Respecting Users

Marshall Van Alstyne & Alisa Lenart

(forthcoming in Communications of the ACM)

Transaction data is like a friendship tie: both parties must respect the relationship and if one party exploits it the relationship sours. As data becomes increasingly valuable, firms must take care not to exploit their users or they will sour their ties. Ethical uses of data cover a spectrum: at one end, using patient data in healthcare to cure patients is little cause for concern. At the other end, selling data to third parties who exploit users is serious cause for concern.[1] Between these two extremes lies a vast grey area where firms need better ways to frame data risks and rewards in order to make better legal and ethical choices. This short essay provides a simple framework and three ways to respectfully improve data use.

Protecting Trust

Trust is a business asset. If you borrow against it, you can quickly become overdrawn. Earning consumer trust requires you to consider:

- Will you secure users' data?
- Will your product/service be reliable?
- Will you protect users' legal rights?
- Will you act ethically?

Users' legal rights include privacy, confidentiality, intellectual property, and contract details found in the terms of service. Laws governing these rights are fact-specific, vary by geography, and often in flux. Yet, even if the law permits you to use data in certain ways, should you? Ethical misuses, which may be legal uses, are often hidden from users and difficult to police. When three media outlets simultaneously reported Facebook's ethical missteps, the Cambridge Analytica scandal stripped more than $100B from Facebook's value.[2]

Risk reduction framework

One simple way to reduce data risk is to take the customer's perspective. Reducing risk means asking:

- Will data use meet the customer's expectations?
- Will they receive fair value in exchange for their data?
- Will they understand how their data is used?
- Will they have choice and control even after their data has been sold?
- How would a firm's use of customer data play out in the court of public opinion; does this differ by country or culture?

Using the customer's perspective to place use-of-data cases on a heat map of reward-versus-risk suggests ethical considerations look like Figure 1.

---

[1] Cadwalladr, C. " 'I made Steve Bannon's psychological warfare tool' Meet the Data War Whistleblower" The Guardian. March 18, 2018.

[2] https://en.wikipedia.org/wiki/Facebook%E2%80%93Cambridge_Analytica_data_scandal

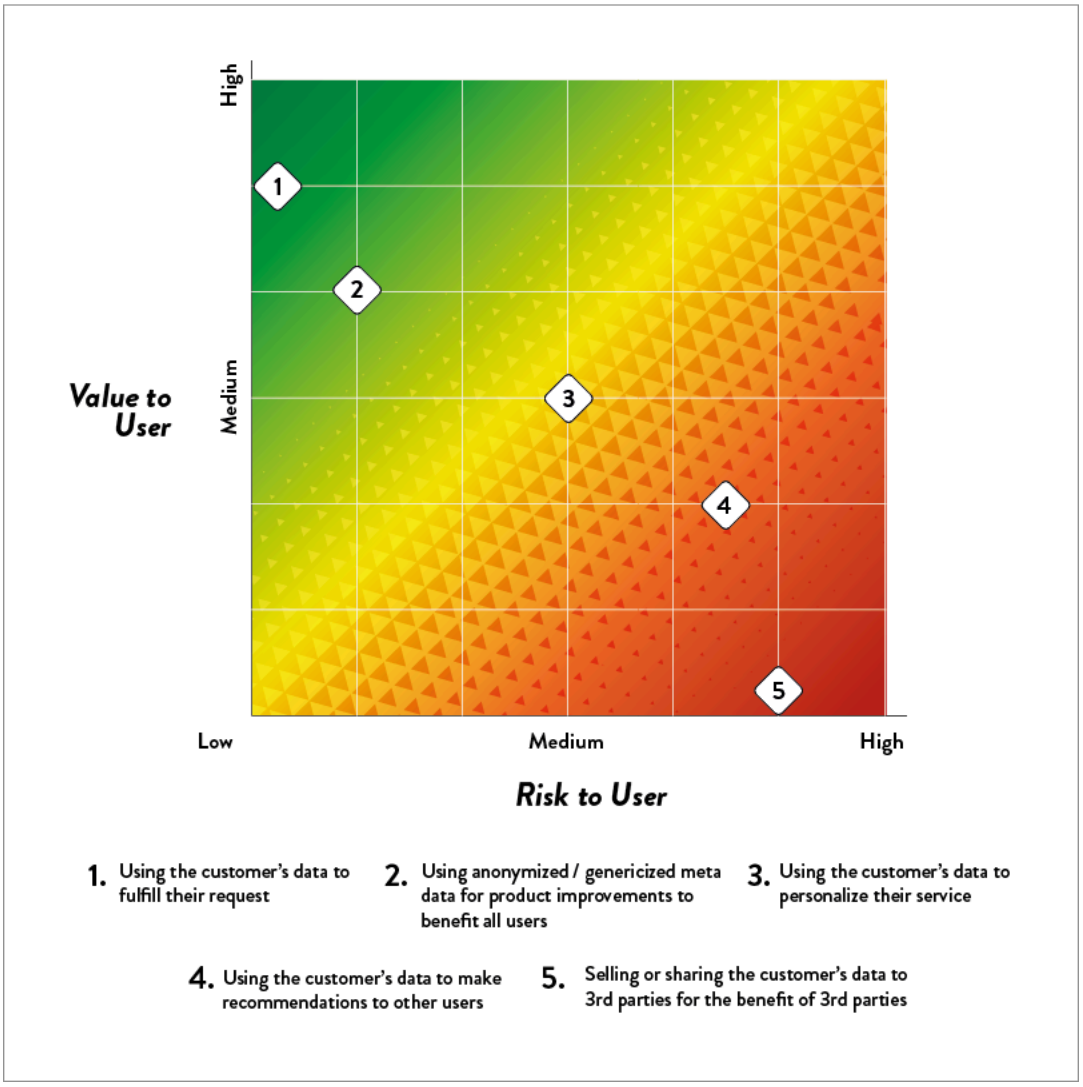

Figure 1 – North-West choices are generally safer than South-East choices.

Evaluation starts from the *perspective of the customer who provides data* - not the business who collects it, nor other users, and certainly not third parties. Ethical and legal risks rise as perspective shifts away from the user.

Risk also rises as data use shifts from a primary to a secondary purpose. A "primary" purpose is that for which the customer originally provided data. A "secondary" purpose is using the data for something else. Pregnancy apps are a great example. They collect extremely personal data and use it to deliver high-stakes insights. Providing a user with predictions on the days they might ovulate would be a primary purpose.

Packaging their data with that of co-workers and selling it to employers or insurance companies to project maternity costs would be a secondary purpose.[3]

The more personal the data, the greater the risk to the firm and the consumer. In the green low-risk quadrant, anonymized data could deliver value to all users. A music streaming app might analyze the size of customer files and speed at which they travel to improve performance for everyone. Risk rises if analysis touches content the customer considers confidential or when one firm fears data leakage to competitors. For example, competing services Spotify and Pandora might contract with the same cloud provider, who mines their content for analytic insights. A problem then arises if Pandora gets insights from Spotify data. To maintain trust and reduce risk, data analysis must give each data source full transparency about how a service works together with a compelling value proposition.

Given this framework, here are three ways to improve the reward-to-risk ratio.

1) Design for Value, Share with Users

Designing for user value expresses the obvious rubric: create more benefit than cost. Users are more or less willing to share data based on whether you give or take value. The same person might happily share a resume that leads to a job opportunity but actively withhold that resume if it were used for psychographic profiling and voter manipulation. Willingness to share data depends on *how* it is used and *who* gets the benefits. The 'how' should be ethical and the 'who' should emphasize the sharer. Design enters this calculation as it affects both parameters. One story from a grocer and one from an advertiser illustrate the shift in mindset from third party to data source.

Groceries are a low-margin business, leading most grocers to sell customer loyalty data to third parties or use it for price discrimination. This creates little customer value and identifies the most price sensitive buyers. To address this challenge, one brand loyalty expert proposed a solution for a New England grocer. The new policy would use loyalty data to *protect* consumers. It would identify products with sugar, MSG, gluten, and peanuts and flag these on behalf of diabetics, celiacs, and people allergic to peanuts. This would decrease sales on flagged products and anger certain distributors. But, as a consumer, imagine your loyalty to a grocer who protects you from bloating, nausea, or diarrhea. Is it worth a price premium to be actively protected from harm? Under a protect-the-user policy, consumers may actively volunteer information to receive this value. Protecting customers increases both their willingness to participate and their willingness to pay. It shifts a grocer from low margins to loyal sales.

A second story concerns a ratings agency that tracks TV ad views to help networks price advertising. Concerned that viewers were skipping ads, the ratings agency designed ad-tracking and motion-sensing technology to learn what viewers saw at each instant. However, it was tone deaf to customer value. Even when paid, few viewers wanted spy systems in their homes just so third parties could learn about their private lives and sell ads.[4] A redesign focused on a mutually beneficial relationship. First, users gained control and could turn the system off. Second, repurposed motion sensors provided free home security

[3] https://www.washingtonpost.com/technology/2019/04/10/tracking-your-pregnancy-an-app-may-be-more-public-than-you-think/?arc404=true

[4] A competitor that did this without consent got sued: https://www.classlawgroup.com/vizio-smart-tv-privacy-lawsuit/ .

and fire protection. These features compared favorably to less sophisticated systems that cost over $30 per month. Although not yet fully deployed, a more sophisticated version could track "senior moments" and help trace likely locations of mislaid keys, glasses, and phones. Third, dashboards let users see their habits as well as any TV network could and manage the results. User-centered design provided transparency, choice, control, and fair value exchange. Ironically, J. Edgar Hoover used FBI spy systems to develop secret citizen files and harass political activists leading to public outcry in the 1950s and 60s,[5] yet today Amazon and Google have sold more than 98 million home-listening devices in exchange for data on sports, news, weather, and users' personal calendars.[6]

2) Save the Data, Discard the Detail

A second approach balances analytic flexibility with privacy. This method hinges on the insight that delivering value from data need not require access to *raw* data. Masked data, which cannot be converted back to its original form or linked to its source, can still permit analysis and even allow researchers to later ask unanticipated questions. Masked content goes beyond masked identity.

One such algorithm works by balancing two competing properties. The first step transmutes and reduces total available data; the second step aggregates sources.

The first step represents lossy compression, where inessential entropy is discarded. Hashing represents one example. In the case of text, this step systematically makes individual words difficult to reconstruct by using morphological properties of language to shed linguistic detail while retaining root structure. It also discards enough information that subverting the algorithm via cryptanalysis becomes difficult.

The second step bundles masked information across individuals or across time in order to supply a corpus large enough to provide statistically-meaningful pattern analysis.

A more aggressive first stage provides greater privacy. A more aggressive second stage provides greater confidence in data analysis. To add protection, use lossier compression. To recover statistical power, aggregate more samples.[7] Individuals and individual messages get harder to read but populations and patterns get easier to resolve.[8]

Researchers used this method to analyze the relationships among email habits, content, and productivity of white-collar workers; yet no researcher could read any email involved in the study. Managers wanted to know, for example, 'Does social network centrality predict productivity?' – yes. 'Is communications diversity associated with productivity?' – yes, but with an inverted-U shape. More content diversity

---

[5] Theoharis, A. G., & Cox, J. S. (1988). *The boss: J. Edgar Hoover and the great American inquisition*. Philadelphia: Temple University Press.

[6] Cumulative sales since 2016. Source: https://voicebot.ai/2019/06/18/loup-ventures-says-75-of-u-s-households-will-have-smart-speakers-by-2025-google-to-surpass-amazon-in-market-share/

[7] There are key tradeoffs. See: Li, N., Li, T., & Venkatasubramanian, S. (2007, April). t-closeness: Privacy beyond k-anonymity and l-diversity. *IEEE 23rd International Conference on Data Engineering* (pp. 106-115). IEEE.

[8] Reynolds, M., Van Alstyne, M. Aral, S. (2009) "Privacy Preservation of Measurement Functions on Hashed Text" Annual Security Conference. "Discourses in Security Assurance & Privacy," Las Vegas, NV. April 15-16, 2009: Information Institute Publishing. p 41-45.

predicts revenues up to a point past which it implies lack of focus.[9] Using this technique, one could ask new questions to understand information diffusion, network diversity, responsiveness, content overlap, or even ad word targeting without reading literal content. Analysis of masked geolocation data or numbers could proceed analogously.

Of course, data masking must avoid infringing intellectual property rights and protect users' other legal rights but keeping only masked data has three major benefits. It boosts willingness to share data. It reduces recording bias from users modifying their behavior. Most importantly, it reduces users' risks even in cases of firms complying with the process of legal discovery or suffering a data breach.

3) Save the Algorithm, Discard the Data:

A third approach uses any number of machine-learning algorithms – neural networks, regression, random forests, k-means clustering, naïve Bayes, etc. – to build a model of the world; then it saves that model but discards the data. Using this method, no data exists that could later be breached, compromised, de-anonymized, sold, or stolen yet it remains possible to classify a new image or to predict a new product's popularity. Another method, secure multiparty computation (MPC), splits the data among several independent parties. Each party can perform calculations on their partition but not see how the results combine. A third party combines results but cannot see the data. This limits access to data during the same calculation whereas discarding data limits access in future calculations.[10]

An advantage of saving the model and discarding the data is that training on complete data can create models with great accuracy. The AlphaGo machine learning algorithm beat the world's expert at the game of GO.[11] A different algorithm beat human lawyers at analyzing risks present in non-disclosure agreements (NDAs).[12] A third algorithm predicts the onset of strokes and heart attacks more accurately than doctors.[13] Another detects breast cancers with 99% accuracy.[14] The disadvantage of finely-tuned machine-learning models is that they cannot be used for purposes outside their training. You cannot get good answers to questions you did not ask. If raw data are gone, there is no re-training option. By contrast, the advantage of saving masked data as in point 2 above is that one can ask new questions that one overlooked initially. However, the disadvantage is that the loss of information causes model accuracy to fall relative to analysis of raw data.

Keeping only the final trained algorithm naturally limits future applications to a primary purpose – the one used to train the model. Using the model for a different purpose would require access to raw data for retraining. The absence of this data limits secondary uses, which limits legal and ethical risk.

---

[9] Aral, S., & Van Alstyne, M. (2011). The diversity-bandwidth trade-off. *American Journal of Sociology*, 117(1), 90-171.

[10] See https://mc.ai/secure-machine-learning-research-with-crypten/

[11] https://en.wikipedia.org/wiki/AlphaGo

[12] https://www.techspot.com/news/77189-machine-learning-algorithm-beats-20-lawyers-nda-legal.html

[13] https://spectrum.ieee.org/the-human-os/biomedical/diagnostics/ai-predicts-heart-attacks-more-accurately-than-standard-doctor-method

[14] https://healthitanalytics.com/news/google-deep-learning-tool-99-accurate-at-breast-cancer-detection

Conclusion

These three approaches – designing for user benefit, saving masked data, and saving masked algorithms – each improve a user's reward-to-risk ratio. Design for user benefit increases the value to users and pushes points North on the Figure 1 heat map. Saving masked data and masked algorithms reduces user profiling, secondary uses, and third-party access, pushing points West in Figure 1. Together, these three approaches offer a range of ways to deliver value from data analysis while protecting users and respecting their trust. Approaching data analysis from the perspective of the user who provided data is not only good business and legal advice but also a way to strengthen ethics and relations with users.