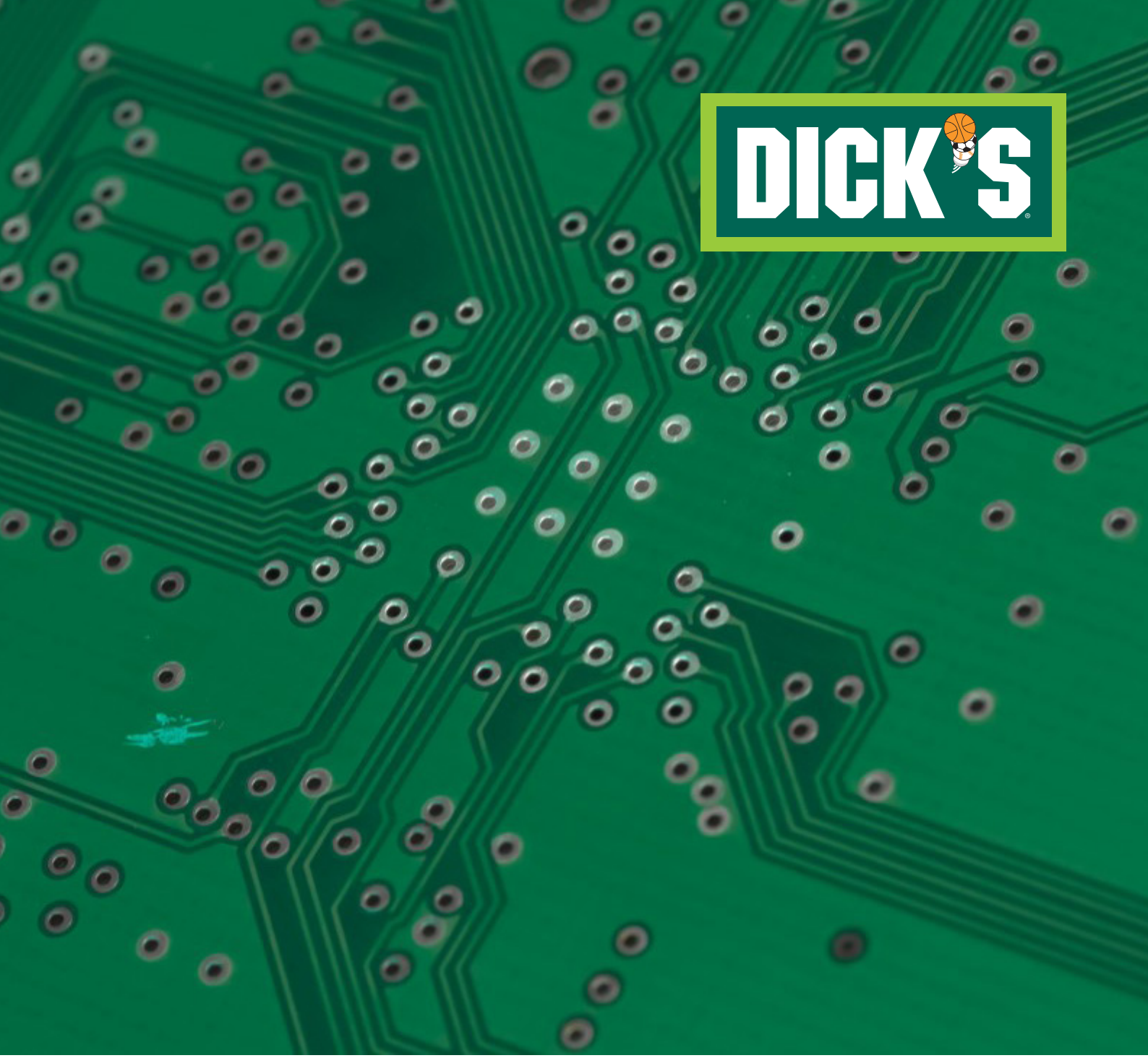


Framework for Responsible Data Science at Dick's Sporting Goods

A principle-based tool designed by Responsible Data Science at Pitt to guide dimensions of data work at Dick's Sporting Goods.

Responsible Data Science at Pitt

Overview

Defining Responsible Data Science (RDS)

Responsible Data Science (RDS) is the intentional alignment of the planning, development, application, and improvement of data-based and computational tools with the values of communities and organizations to empower positive decisions and mitigate harms.

Responsible Data Science at Dick's Sporting Goods

RDS principles fit into the greater core values of Dick's Sporting Goods: dedication, optimism, integrity, and authenticity. By understanding the role of RDS principles within these core values, we can establish a shared language for discussing the role of responsible data science at DSG.

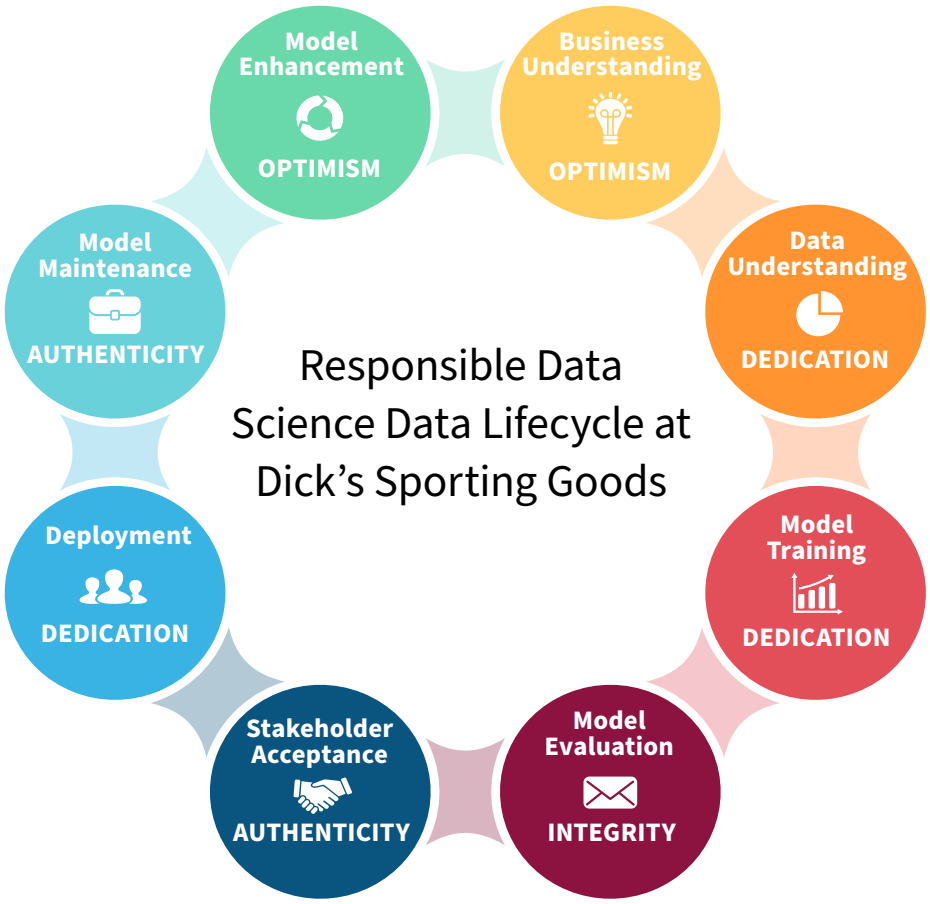
Dedication
- Data Quality
- Security
- Accountability
- Adaptability

Optimism
- Inclusivity
- Explainability
- Collaboration
- Beneficence

Integrity
- Fairness
- Privacy
- Ethics
- Transparency

Authenticity
- Sustainability
- Proportionality
- Accessibility
- Impact Assessment

DSG core values, and the RDS values they encompass, come into play at every stage of the data life cycle. Below is a spotlight on the DSG value that is most relevant to responsible data science at each stage.



1. Understanding Principles of Responsible Data Science in Context

Responsible Data Science (RDS) is the intentional alignment of the planning, development, application, and improvement of data-based and computational tools with the values of communities and organizations to empower positive decisions and mitigate harms. Principle-based frameworks are guidelines that support the examination of data science practices. The RDS principles discussed below are are the basis of responsible data work, and the product of interdisciplinary consensus-building about what RDS means and looks like.

RDS principles fit into the greater core values of Dick's Sporting Goods: dedication, optimism, integrity, and authenticity. By understanding the role of RDS principles within these core values, we can establish a shared language for discussing the role of data science at DSG.

Dedication

Commitment to excellence in work, product quality, and investor return.

- **Data Quality:** Ensure accuracy, completeness, and reliability of data.
- **Security:** Protect data and systems from unauthorized access and cyber threats.
- **Accountability:** Assign clear responsibility for data collection and usage.
- **Adaptability:** Continuously refine practices to adapt to evolving challenges, regulations, and technological advancements.

Optimism

Approaching work with a positive mindset, fostering collaboration, and seeing the best in people.

- **Inclusivity:** Engage diverse perspectives and communities in the design and application of data-driven systems.
- **Explainability:** Ensure data decisions are understandable to all stakeholders.
- **Collaboration:** Foster interdisciplinary collaboration for better outcomes.
- **Beneficence:** Do no harm.

Integrity

Upholding legal standards, ensuring fairness, and balancing business requirements with ethical responsibility.

- **Fairness:** Ensure data-driven decisions are free from bias and promote equity.
- **Privacy:** Safeguard individuals' personal and sensitive information.
- **Ethics:** Align data practices with ethical standards and societal values.
- **Transparency:** Maintain openness about data sources and methodologies.

Authenticity

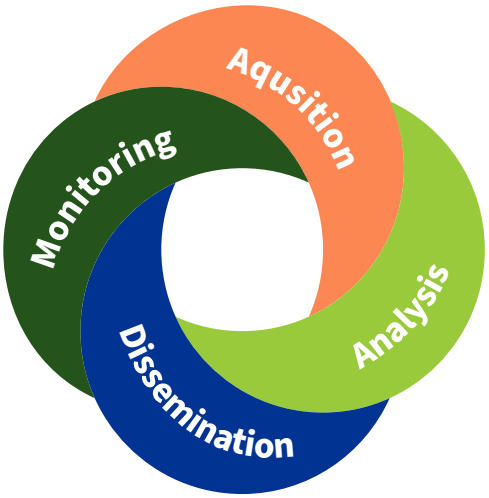
Demonstrating genuine commitment to communities and always striving for the best outcomes for people.

- **Sustainability:** Promote long-term, environmentally-conscious data practices.
- **Proportionality:** Balance the benefits and risks of data science applications.
- **Accessibility:** Make data tools and outcomes available to a broad audience.
- **Impact Assessment:** Evaluate the short and long-term societal impacts of data science projects (cultural, social, and contextual implications).

2. Principle-Aligned Practices for the Life Cycle of Data at Dick's Sporting Goods

The data lifecycle refers to the different stages a unit of data undergoes, from initial collection to when it's no longer considered useful and deleted. It's a continuous process where each phase informs the next. Different decisionmaking points arise throughout the data life cycle that inform the intent, use, form, and control of the data. By applying DSG's core values—which align with RDS's key principles—to our decision-making, we actively promote responsible data science.

The data life cycle has four stages: acquisition, processing and analysis, dissemination, and monitoring. For each stage in the life cycle, we will define the stage, discuss the intersections of DSG and RDS principles, and explore how RDS can impact KPIs and error metrics.



Acquisition

Business Understanding

Sometimes, business and data talk are like different languages. Business understanding refers to the ability to grasp the strategic questions or goals behind model creation. It informs the entire data process. Model results are not the only important insight; this stage can uncover knowledge that enhances future decision-making.

Role of DSG Core Values: OPTIMISM

Responsible data science at this stage requires:

- **Inclusivity,** by ensuring that data models account for different needs and skillsets. This may include outreach to subject matter experts and stakeholders for clarification.
- **Explanability,** by providing a shared language and foundation about the goals and results of data insights. Explanability requires finding common terms between teams.
- **Collaboration,** by bridging the gap between different technical teams and fostering shared ownership and decision-making.
- **Beneficence,** by ensuring data models are designed to serve the best interests of DSG and the communities it supports.

Impact of RDS on KPIs & Error

For data scientists to make meaningful impacts on business strategies, they must understand the goals from a business perspective. Inclusivity and collaboration across teams can guide which KPIs are most important to maximize, as well as which features can help models reduce error.

Data Understanding

Knowing what data is needed to produce a successful model can take trial and error. Data understanding requires exploring of what data exists and where, and how to harness new data sources.

Role of DSG Core Values: DEDICATION

Responsible data science at this stage requires:

- **Data quality,** by identifying inconsistencies or gaps in the data, ensuring that models are built on clean, effective data sources.
- **Security,** by recognizing sensitive data and applying appropriate protections to prevent unauthorized access or misuse.
- **Accountability,** by documenting data sources, assumptions, and limitations and ensuring traceability throughout the modeling process. This includes engaging data engineering partners for awareness when setting up new data processes.
- **Adaptability,** by providing data definitions and resource information, enabling future data scientists to easily interpret and leverage new data.

Impact of RDS on KPIs & Error

Responsibly deciding which data is most high-quality and relevant for the goal at hand helps to reduce noise and error in model predictions. Responsible feature selection also allows stakeholders to understand and trust model decisions.

Analysis

Model Training

At this stage, it is often best to train a baseline model, then move onto more complex modeling approaches. Fine tuning hyperparameters should be done on a few best-performing models.

Role of DSG Core Values: DEDICATION

Responsible data science at this stage requires:

- **Data quality,** by harnessing diverse and unbiased data across demographics.
- **Security,** by ensuring models using sensitive data are protected from unauthorized access.
- **Accountability,** by clearly communicating responsibilities to ensure efficient training pipelines.
- **Adaptability,** by staying up to date on models and optimization techniques revealed by machine learning research.

Impact of RDS on KPIs & Error

Choose models with the most relevent complexity, objective, and efficiency. Complex models are more likely to overfit data, resulting in inaccurate predictions for underrepresented groups.

Model Evaluation

Evaluating model performance is extremely important. Data and buisness teams should work together to ensure that models align with goals.

Role of DSG Core Values: INTEGRITY

Responsible data science at this stage requires:

- **Fairness,** by ensuring models generalize well to diverse data sets and are not biased towards specific groups of people.
- **Privacy,** by safeguarding sensitive information in model evaluation, especially when dealing with personally identifiable or other sensitive information.
- **Ethics,** by evaluating models not only on their performance on business goals, but their ability to uphold ethical standards.
- **Transparency,** by ensuring the model is accessible. Credit should be given to inventors of models and evaluation methods.

Impact of RDS on KPIs & Error

Minority groups are most affected by overfitting inaccuracies from unbalanced datasets. Ensuring fair evaluation across demographics is important for minimizing prediction error and maximizng KPIs.

Dissemination

Stakeholder Acceptance

This stage is about discussing model results, agreeing to a testing plan, and securing the final buy-in on the model with stakeholders.

Role of DSG Core Values: AUTHENTICITY

Responsible data science at this stage requires:

- **Sustainability,** by considering the maintenance and resource needs of the model. Explaining how long it takes to make and improve models to stakeholders may be needed.
- **Proportionality,** by balancing the benefits and risks of the the model. Address stakeholder misconceptions if they over or underestimate what tools are capable of.
- **Accessibility,** by presenting results in an actionable, outcome-oriented manner and avoiding dwelling on complex methods.
- **Impact Assessment,** by evaluating the short and long-term impacts of data science projects.

Impact of RDS on KPIs & Error

Discussions with stakeholders and experts in the field of the model are essential for ensuring that models are maximizing KPIs to the extent that the business expects. These experts can also provide essential knowledge related to the task of the model which can lead to lower model error.

Deployment

Deployment involves building the necessary structures to launch a model into production, like testing and outlining the pipeline for the model to dev and incorporating a CI/CD process.

Role of DSG Core Values: DEDICATION

Responsible data science at this stage requires:

- **Data quality,** by ensuring the quality and accuracy of all tests en route to deployment (both unit tests and integration tests).
- **Security,** by ensuring that models deployed to production do not leak sensitive information to unauthorized users.
- **Accountability,** by ensuring all parties understand their responsibilities during collaboration between data scientists and machine learning engineers.
- **Adaptability,** as data scientists dynamically respond to evolving challenges and regulations.

Impact of RDS on KPIs & Error

Transparency about what data is collected on end-users may result in users providing more accurate data, leading to more effective models. Inclusivity is also important when end-users come from diverse demographics. Ensuring that any user can effectively make use of data science tools will ensure that KPIs are maximized across various groups.

Monitoring

Model Maintenance

During a model's lifetime, data or the model itself can drift as trends in the data change. It is important that models are maintained to uphold their performance as statistical properties of data change.

Role of DSG Core Values: AUTHENTICITY

Responsible data science at this stage requires:

- **Sustainability,** by taking steps to mitigate data drift and maintain consistent model performance over time.
- **Proportionality,** by carefully weighing the pros and cons of changes before re-deploying models.
- **Accessibility,** by making model modifications clear and understandable to users.
- **Impact Assessment,** by considering both the short- and long-term societal impacts and adjusting models to prevent malicious effects.

Impact of RDS on KPIs & Error

Model drift can occur when relationships between input features and target variables change over time. The values of optimism and authenticity should be followed by ensuring that models are trained on up-to-date data. Additionally, model biases can accumulate over time, which could unfairly affect different demographics. It is important that data scientists are dedicated to ensuring high quality data when maintaining existing models to uphold their accuracy and continue maximizing KPIs.

Model Enhancements

During the data science process, data scientists may realize the importance of new features or learn about new methods. These features and methods can be added to models in order to facilitate better performance.

Role of DSG Core Values: OPTIMISM

Responsible data science at this stage requires:

- **Inclusivity,** by engaging with diverse perspectives and communities to enhance model generalizability when biases toward specific demographics are identified.
- **Explainability,** by ensuring that model enhancements are easily understandable to stakeholders and aligned with strategic business goals.
- **Collaboration,** by fostering interdisciplinary and cross-sector partnerships to generate new ideas and incorporate informed opinions on feature selection.
- **Beneficence,** by ensuring that model enhancements do not cause harm to people, societies, or the environment, even when unintended consequences arise.

Impact of RDS on KPIs & Error

RDS principles such as inclusivity and explainability ensure that all communities are considered when enhancing models, resulting in models that are applicable to diverse data sets. This generalization helps reduce error metrics in the models. Explainability is also extremely important in data analysis since stakeholders will be more willing to trust analyses to increase KPIs.

3. Case Study Example

Example: Virtual AI assistant to help customers navigate the Dicks Sporting Goods website.

Acquisition

Buisness Understanding

Data scientists are in contact with product teams to ensure the correct products are being marketed and that the right website features are being developed, and collaborate with marketing and customer service teams to understand the current landscape and ensure business goal sare targeted.

Data Understanding

Data scientists garner a good grasp of the kinds of data necessary to pursue an AI assistant; they ensure this data comprehensive, well-labeled, and representative of the target use cases (based on buisness understanding) to minimize bias and improve model performance. Data scientists also ensure that sensitive consumer data is securely stored.

Analysis

Model Training

Data science team members split duties for training the model to promote accountability within the team. They also stay up to date with methods for training models. This could involve understanding new APIs for LLMs.

Model Evaluation

Data scientists ensure that the model responds appropriately to all demographics. This could include understanding diverse dialects of language. They also confirm that the model does not hallucinate if it cannot find a response, or respond in a way that may have negative impacts on users.

Dissemination

Stakeholder Acceptance

Strengths and weaknesses of the assistant are clearly communicated to stakeholders. When a colleague states their concern that the assistant will impact their job, data scientists explain how AI at DSG will only automate basic and repetitive tasks, allowing employees to focus on higher-level, strategic work that requires human judgment and creativity. Data scientists work with stakeholders to evaluate the sustainability of the model as well as its impact on society.

Deployment

Multiple teams work together to get the assistant into production. Keeping each team accountable for their responsibilities ensures a swift and efficient transition to production. Data scientists rigorously test their pipeline and model for any security leaks, ensuring that sensitive consumer data that may be provided to the assistant remains confidential.

Monitoring

Model Maintenance

As language changes and DSG offers new products and services, data scientists make sure the assistance is able to understand new slang and terms. Maintenance could also involve addition of new protocols for ordering. Data scientists collaborate with the necessary teams to ensure new policies are added to the model appropriately.

Model Enhancements

New LLM API features are added regularly. Data scientists should ensure that these features are explainable and justifiable to stakeholders to make certain that they will promote the business goals. Data scientists should also confirm that any enhancements made to the assistant do not disproportionately affect certain groups of users.