Data Science Ethics

Proposing an ethics framework for data scientists.

Eliza R. Starr

5 April 2019

Contents

1. Executive Summary

1.1 Introduction

Interest in data ethics is growing because companies, governments, and users are more aware of the unintended consequences of capturing and using personal data. The creation of new laws and a rise in awareness is forcing the Wild West of data science to civilize. Companies and their data scientists must anticipate these changes in order to reduce legal and social risks. Because the ethical responsibility of collecting and using data is increasing, businesses need to explore new ways of ensuring the safety of users and data.

This paper focuses on data science processes rather than data governance, security, and compliance processes. Data science ethics is a relatively new topic for companies to address, and the completion of the following tasks will help them understand its ethical responsibilities and minimize legal and social risks in the data science relm.

1) Explore the recent history and discussions that motivate the need for a data ethics framework.
2) Introduce existing data ethics solutions proposed by industry professionals.
3) Choose and implement a data ethics solution.

1.2 Literature Review

The Facebook-Cambridge Analytica data scandal improved the reception of government policies like the General Data Protection Regulation (GDPR) and its American counterpart, the California Consumer Privacy Act. These events and laws drew government and public attention to data ethics issues. Ethics concerns can be focused further onto the work of data scientists. The field of data science is especially prone to ethical uncertainty and therefore data science ethics must be addressed by all industries.

1.3 Methodology

The solutions proposed by this paper are sourced from various articles, opinion pieces, and ethical frameworks. The optimal solution is then chosen and implemented in Jira, an issue and project tracking software.

1.4 Solutions

The following solutions exist for companies seeking to be responsible with their data.

1) Train employees in data ethics.
2) Create or adopt a code of data ethics.
3) Encourage employees to flag projects for ethical evaluation.
4) Use bias-checking software for machine learning models.
5) Report misconduct on reputation-influencing websites like Glassdoor and Better Business Bureau.
6) Require the completion of an ethics workbook at the start of new data-driven projects.

1.5 Recommendation

If a company does not yet have standardized ethical processes in place, a simple and enforceable solution is to integrate an ethics workbook into the project tracking software already used by the company. This is done by adding custom fields to issue tickets and adding custom pipelines.

1.6 Conclusions

Data ethics is a hot topic in all industries that make use of data. Many strategies exist that can decrease social and legal risk of a company using data.

2. Introduction

According to Issak, "[t]he discovery that Facebook gave unfettered and unauthorized access to personally identifiable information (PII) of more than 87 million unsuspecting Facebook users to the data firm Cambridge Analytica has fueled growing interest in the debate over technology's societal impact and risks to citizens' privacy and well-being" (2018). Governments, companies, not-for-profit organizations, and data practitioners have since contributed to an ongoing dialogue about the risks of and solutions to the use of private data by companies.

The purposes of this report are to learn about the global state of data ethics, to identify ethics solutions, and to recommend an additional ethical precaution that a company could implement in the future. Therefore, the report is separated into the following three parts.

1) Present research on the recent history and conversation around data science ethics. This information will help companies assess the urgency of implementing data science ethics standards.
2) Introduce existing data ethics solutions proposed by governments, not-for-profit organizations, and industry professionals.
3) Standardize a data ethics workbook to be used by data scientists and analysts at the start of every project. This workbook would reduce the social and legal risk companies take when using their data by recording the intentions, limitations, and impacts of each project.

3. Literature Review

The Facebook-Cambridge Analytica scandal, General Data Protection Regulation (GDPR), and California Consumer Privacy Act motivate this project because they dictate which data practices society is uncomfortable with and shows where the conversation about data ethics is moving.

Cambridge Analytica obtained Facebook user data for use in an approved project, but then extended the data for use in a project having to do with the United States' 2016 presidential elections (Issak 2018). The scandal generalizes to the following issues: some users did not consent to data collection, the use case for the data was unapproved, and the data was unnecessarily identifiable.

Troiani (2018) summarizes the GDPR into five rights: required consent for data collection, 72 hour security breach notification, know how personal data is used, data erased, obtain and use personal data. According to Simberkoff (2018) the GDPR was years in the making before the data scandal broke, but the news accelerated investigations and acts of enforcement carried out by European Data Protection Authorities under the regulation.

The GDPR likely inspired the California Consumer Privacy Act (Assembly Bill No. 375, 2018), which will take effect in January 2020. This act grants consumers rights similar to those in the GDPR. For example, the right requiring “California-based organizations to obtain explicit consent from users before sharing or selling their data with vendors, providers and partners” (Simberkoff, 2018). Simberkoff believes that the GDPR and California Consumer Privacy Act “moved the world more closely to a regulatory environment where companies must do the right thing and actively work to change privacy laws to further protect consumers” (Simberkoff, 2018).

The GDPR protects the data of all EU citizens and residents even if the data is collected or stored by companies outside of the EU (Troiani, 2018). Theoretically, any US based company must comply with the regulation if the company knowingly or unknowingly collects any data on an EU citizen or resident. The real estate industry is no exception. Real estate companies are advised to take inventory of their data, add a consent form to websites, and discuss how the company will respond to users exercising their rights (NAR Legal Affairs, 2018).

Touw (2018) predicts that more regulations like the GDPR will be enforced in the future. These regulations will not only dictate what data is collected but also how data is used. This attention to detail means that ethical responsibilities are cascading to data scientists who use data to create product features and business insight.

Data analytics can be misused and, therefore, require rules and supervision (Gupta, 2015). Data analytics is not an exact science, its problem solving methodologies have not been standardized, and it often requires data scientists to use their best judgement and introduce human bias. Touw categorizes big data responsibilities into access restrictions, purpose restrictions, and model monitoring. Firstly, data should only be accessible to employees who need it, rather to the entire company. Then, the use of data should be restricted to clearly defined and approved purposes. Lastly, deployed models should be monitored for the quality of their input and output. Woodie (2017) adds technical detail to this last responsibility. He recognizes that because data is inherently biased, the machine learning models used to predict from data will be biased as well. Companies must reduce and monitor the impact of this model bias. The Solutions section will present ways of enforcing the responsibilities Touw suggests.

4. Methodology

Various articles, journal papers, opinion pieces, and ethical frameworks supplied the information presented in the literature review and solutions section. The solution proposed in the recommendation section was chosen according to the cited research and experiential knowledge of how a data science team operates.

5. Solutions

Research presents the five following data science ethics solutions: employee training, codes of ethics, processes for flagging projects for evaluation, bias-checking software, reputation-influencing websites, and ethics workbooks.

One way to improve ethical practices is to train future employees on the topic of ethics. According to Leetaru (2018), universities must incorporate big data ethics into their data science curricula to produce graduates whose ethical principles will influence the industries they work for. Gupta agrees that ethics training should start at the university in order to influence data-driven industries to expect ethical conduct. Johnson (2019) noticed that more universities have incorporated ethics into their data science curriculum by including ethics and humanities courses. These additions introduce students to the holistic and interdisciplinary nature of data science.

Another option is to create or adopt a code of data ethics that defines principles to uphold. This code is comparable to a core value statement, except that the code is more detailed and focuses specifically on data practices. Data for Democracy, a volunteer community, created the Global Data Ethics Project (2018), which pioneers this potential solution. Those who sign this code pledge to act through fairness, openness, reliability, trust, and social benefit in the data science context clearly defined in the document.

Wheeler (2018) does not recommend the adoption of ethical codes and regulations because codes are effectively goal statements that cannot be enforced and regulations invite people to search for loopholes and operate on the arbitrary ethical boundary that is set. He suggests a number of enforceable alternatives including consequence-checking tools, an ethical flagging system, and crowd-sourced reputation platforms. Tools such as Aequitas, created by the University of Chicago (2018), and software that tests for the Simpson's Paradox, like the method created by Alipourfard (2018), will expose unintended biases and patterns that could influence a machine learning model. The flagging system allows employees to flag projects that seem vaguely, specifically, or strongly unethical. The flagged project will then either be casually discussed, reviewed, or put on hold for serious evaluation. Lastly, concerned employees could post their ethical concerns about a company on public reputation sites like Glassdoor and Better Business Bureau. These reputation-impacting actions may then encourage industries to change their behavior.

A company can also require data scientists to complete an ethics workbook at the start of new data-driven projects. The United Kingdom's Department for Digital, Culture, Media and Sport published the Data Ethics Framework, which includes a data science workbook to guide data practices in both the government and public sectors (2018). The workbook has seven sections each containing six to ten questions. Each section focuses on one of the following principles: user and public benefit, legislation and codes, necessary data collection, data limitations, robust practices, transparency, and responsible implementation. Each team can decide how often they refer to the workbook for each project.

6. Recommendation

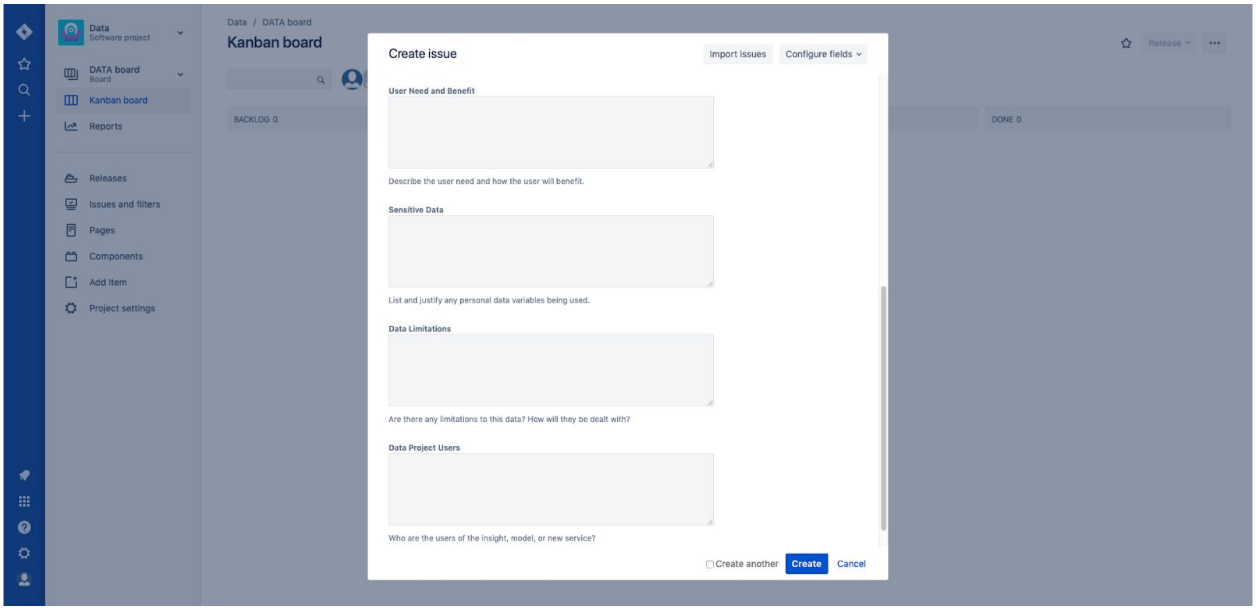
In the near future, governments and users may expect data-driven companies to be more intentional and transparent about their data practices. Companies should adopt new ethical precautions as global data standards change.

In the short term, companies can implement some version of a data science workbook. This workbook would include a list of pre-selected questions that data scientists would then answer about each data science project chosen for development. A workbook like this could be filled out in a spreadsheet, text document, slideshow, integrated development environment, version control system, or issue and project tracking software platform.

A data science workbook is easy to implement in an issue and project tracking software like Jira. All that needs to be done is to add custom fields to the "Create issue" form. For data scientists at, an issue ticket can represent one of two levels of abstraction. A ticket is either a project to be completed a task to be completed within a project. In the latter case, the issue tickets may be assigned to an epic. This epic will then represent the project to which the task was assigned.

The set of ethical questions added to this form can be created by the team or selected from other sources. Figure 1 shows four ethics questions added to a "Create issue" form used by data scientists. The question fields are titled and described as follows.

1) User Need and Benefit: Describe the user need and how the user will benefit.
2) Sensitive Data: List and justify any personal data variables being used.
3) Data Limitations: Are there any limitations to this data? How will they be dealt with?
4) Data Project Users: Who are the users of the insight, model, or new service?



Create issue

User Need and Benefit

Describe the user need and how the user will benefit.

Sensitive Data

List and justify any personal data variables being used.

Data Limitations

Are there any limitations to this data? How will they be dealt with?

Data Project Users

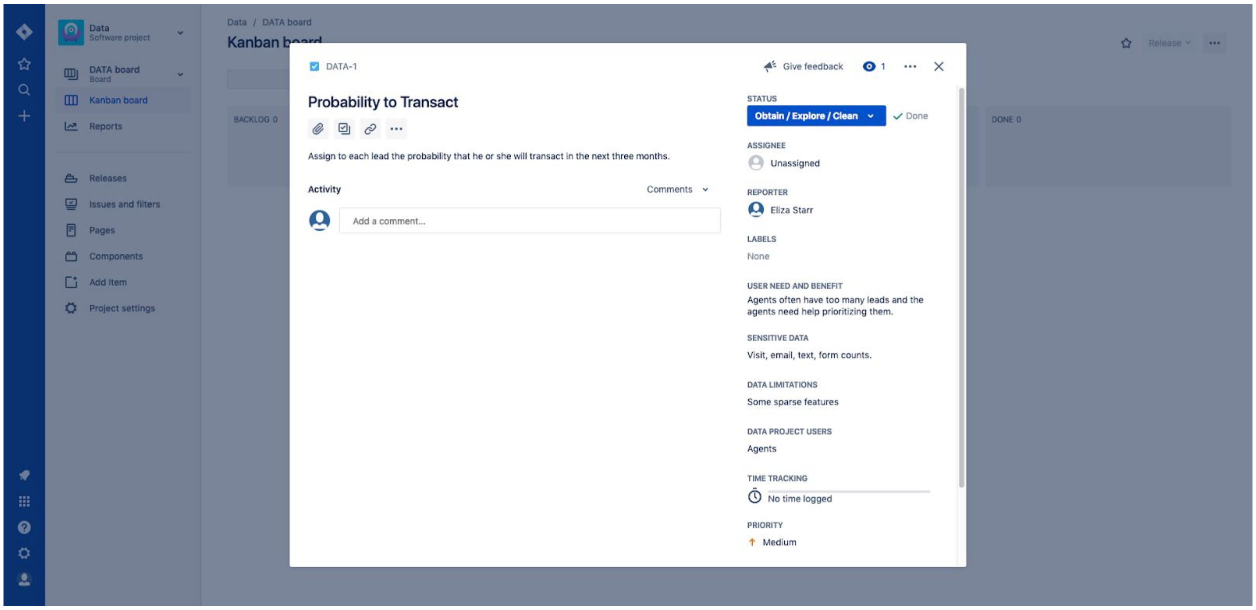
Who are the users of the insight, model, or new service?

Figure 1. The “Create issue” form adapted to be a data science workbook.

Figure 2 shows an example issue ticket created with the new “Create issue” form from Figure 1. This ticket represents a project called Probability to Transact. The right column of the ticket displays the answered ethical questions. The following answers were given.

1) User Need and Benefit: Agents often have too many leads and the agents need help prioritizing them.
2) Sensitive Data: Visit, email, text, form counts
3) Data Limitations: Some sparse features
4) Data Project Users: Agents

These questions can be answered in as little or as much detail as the data scientist wants. This project ticket owner does not go into much detail. This might be because the project is in an early planning phase and not much is known about the project’s goals, users, or data. When the project develops further, the data scientist can expand upon these answers.



DATA-1

Probability to Transact

Assign to each lead the probability that he or she will transact in the next three months.

Activity — Comments

Add a comment...

STATUS
Obtain / Explore / Clean ✓ Done

ASSIGNEE
Unassigned

REPORTER
Eliza Starr

LABELS
None

USER NEED AND BENEFIT
Agents often have too many leads and the agents need help prioritizing them.

SENSITIVE DATA
Visit, email, text, form counts.

DATA LIMITATIONS
Some sparse features

DATA PROJECT USERS
Agents

TIME TRACKING
No time logged

PRIORITY
↑ Medium

Figure 2. An example data science issue with the ethics workbook fields filled out.

The ticket from Figure 2 is shown below in Figure 3. The ticket has been moved from the "Backlog," where newly created issue tickets are placed in this Jira environment, to the "Obtain / Explore / Clean" pipeline.

The workflow organization depicted in Figure 3 is not the main focus of this data science workbook proposal, but it is a topic worth expanding on. A common pipeline configuration includes "Selected for Development," "In Progress," "Done" stages, whereas Figure 3's workflow has the following pipelines.

1) Backlog
2) Obtain / Explore / Clean
3) Model / Interpret
4) Done

This custom workflow represents a common data science project life cycle. A data science project or task is moved to these pipelines according to which stage the project is currently in. First, projects or tasks are added to the backlog, where they wait to be worked on. When an issue ticket is moved to the next pipeline, the data scientist is obtaining, exploring, and cleaning the data needed for the project. Next, the issue ticket can proceed to the "Model / Interpret" pipeline when the data is ready for modeling. In the same stage, the resulting model is interpreted by the data scientist. The issue ticket can then be moved back to an earlier stage or to the "Done" pipeline.

These stages do not encompass the workflow required by all data science projects. For example, a “Deploy” pipeline could be placed between “Model / Interpret” and “Done” if the resulting model is selected for deployment.

A workflow customization like this would give data science work more transparency. Jira can record Pipeline changes so that the progression of data science projects are recorded. This metadata tracks how and when data is being used though each project’s life cycle.

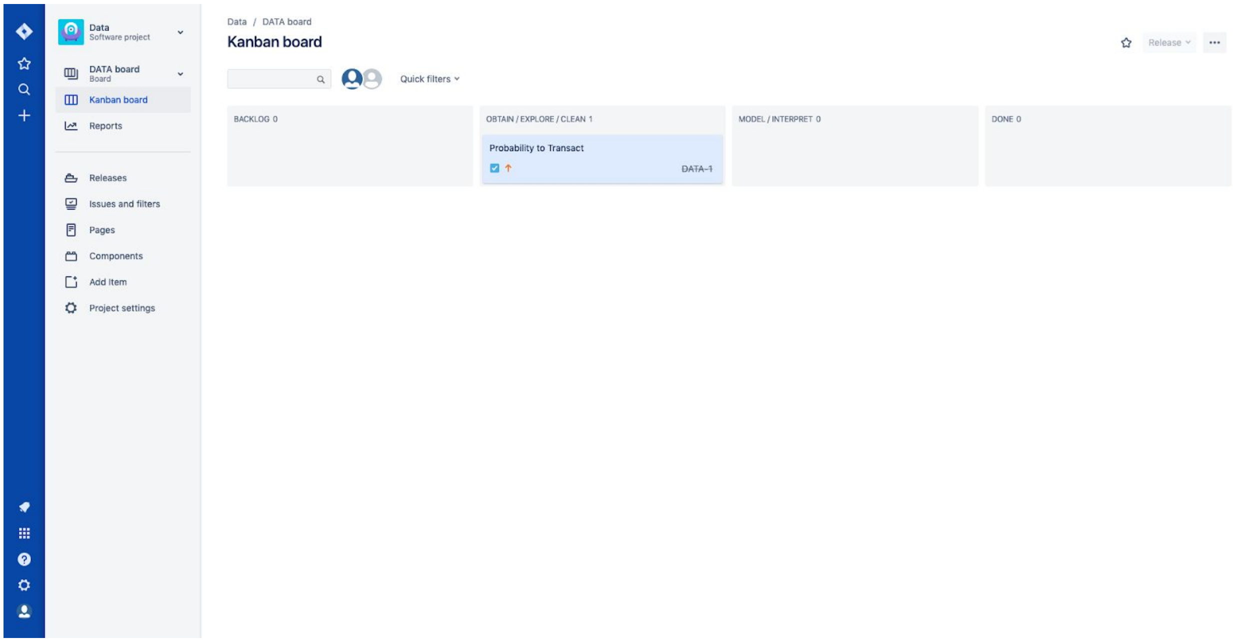


Figure 3. The example issue in the “Obtain / Explore / Clean” stage of the data science workflow.

Figures 4-8 in the Appendix show how the custom fields and workflow were added to Jira.

7. Conclusion

Data ethics are a growing concern of governments and of the public as a result of recent data breaches and scandals. Policies like the General Data Protection Regulation (GDPR) have been drafted and enacted to regulate an increasingly data-driven world. Adopting additional data ethics precautions can only benefit companies in this changing environment.

As companies establish data science teams, they should create data science practices that will instill a sense of ethical responsibility in their growing team. Solutions include ethical training, a code of ethics, a project flagging system, bias-checking software, reputation-influencing websites, and ethics workbooks. Companies could easily create an integrated, customized data ethics workbook in their issue and project tracking software. Proactively introducing this workbook will protect companies and their users from legal and social risks in the future.

8. Bibliography

Alipourfard, N. (2018, December 28). Trend Simpson's Paradox [Computer software]. Retrieved April 2, 2019, from https://github.com/ninoch/Trend-Simpsons-Paradox/

Assembly Bill No. 375. (2018, June 29). Retrieved April 4, 2019, from https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=201720180AB375

Data Ethics Framework. (2018, June 13). Retrieved April 4, 2019, from https://www.gov.uk/government/publications/data-ethics-framework/data-ethics-framework

General Data Protection Regulation. (n.d.). Retrieved April 4, 2019, from https://gdpr-info.eu/

Global Data Ethics Project. (2018, November 25). Retrieved April 5, 2019, from https://www.datafordemocracy.org/project/global-data-ethics-project

Gupta, B. (2015, October). KDnuggets. Retrieved March 14, 2019, from https://www.kdnuggets.com/2015/10/ethics-data-analytics.html

Isaak, J., & Hanna, M. J. (2018). User Data Privacy: Facebook, Cambridge Analytica, and Privacy Protection. IEEE Computer, 51(8), 56-59. doi:10.1109/mc.2018.3191268

Johnson, S. (2019, January 11). University Data Science Programs Turn to Ethics and the Humanities - EdSurge News. Retrieved March 5, 2019, from https://www.edsurge.com/news/2019-01-11-university-data-science-programs-turn-to-ethics-and-the-humanities

Leetaru, K. (2018, October 10). Do We Need To Teach Ethics And Empathy To Data Scientists? Retrieved March 5, 2019, from https://www.forbes.com/sites/kalevleetaru/2018/10/08/do-we-need-to-teach-ethics-and-empathy-to-data-scientists/#3e04fa7312ee

NAR Legal Affairs. (2018, May). General Data Protection Regulation: New EU Data Privacy Law May Affect U.S. Businesses. Retrieved April 4, 2019, from https://www.nar.realtor/legal/general-data-protection-regulation-new-eu-data-privacy-law-may-affect-us-businesses

Simberkoff, D. (2018, August 30). How Facebook's Cambridge Analytica Scandal Impacted the Intersection of Privacy and Regulation. Retrieved April 2, 2019, from https://www.cmswire.com/information-management/how-facebooks-cambridge-analytica-scandal-impacted-the-intersection-of-privacy-and-regulation/

Touw, S. (2018, June 22). Ethical Data Science Is Good Data Science. Retrieved March 5, 2019, from https://www.datanami.com/2018/06/22/ethical-data-science-is-good-data-science/

Troiani, M. L. (2018, June 27). What is GDPR and Why Should It Matter to Real Estate Professionals? Retrieved April 2, 2019, from https://www.nvar.com/realtors/laws-ethics/legal-blog/what-is-gdpr-and-why-should-it-matter-to-real-estate-professionals

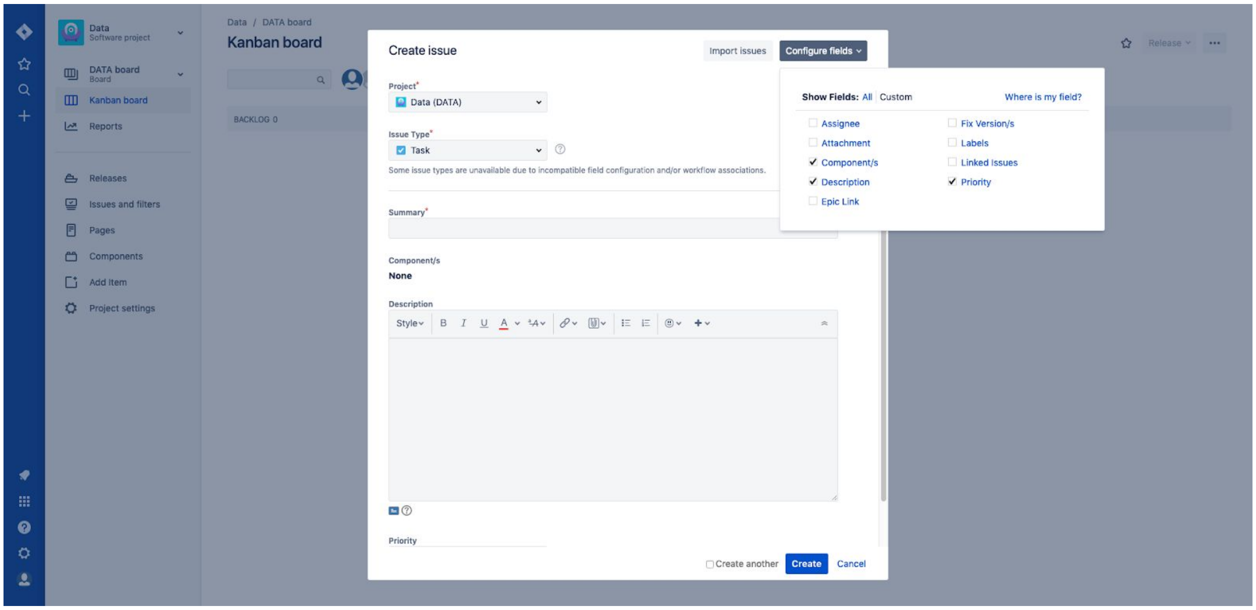
University of Chicago. (2018). Aequitas. Retrieved March 28, 2019, from https://dsapp.uchicago.edu/projects/aequitas/

Wheeler, S. (2018, August 30). Ethical codes vs. ethical code. Retrieved March 5, 2019, from https://towardsdatascience.com/ethical-codes-vs-ethical-code-fea118987a5

Woodie, A. (2017, October 25). Keeping Your Models on the Straight and Narrow. Retrieved March 5, 2019, from https://www.datanami.com/2017/10/24/keeping-models-straight-narrow/

9. Appendix

Jira custom field and workflow creation processes.



Create issue

Project*: Data (DATA)

Issue Type*: Task

Some issue types are unavailable due to incompatible field configuration and/or workflow associations.

Summary*

Component/s: None

Description

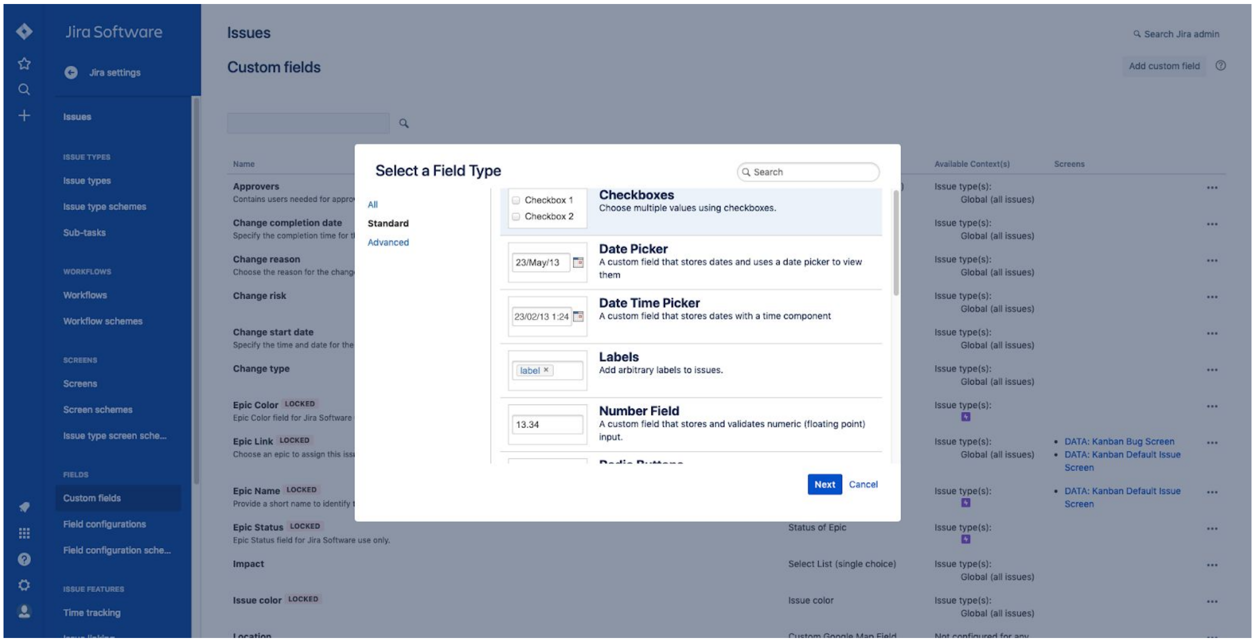
Priority

□ Create another

Show Fields: All | Custom — Where is my field?

□ Assignee	□ Fix Version/s
□ Attachment	□ Labels
✓ Component/s	□ Linked Issues
✓ Description	✓ Priority
□ Epic Link	

Figure 4. The Create Issue prompt before adding custom fields.



Issues

Custom fields

Select a Field Type

Search

All

Standard

Advanced

□ Checkbox 1 □ Checkbox 2 — **Checkboxes**: Choose multiple values using checkboxes.

23/May/13 — **Date Picker**: A custom field that stores dates and uses a date picker to view them

23/02/13 1:24 — **Date Time Picker**: A custom field that stores dates with a time component

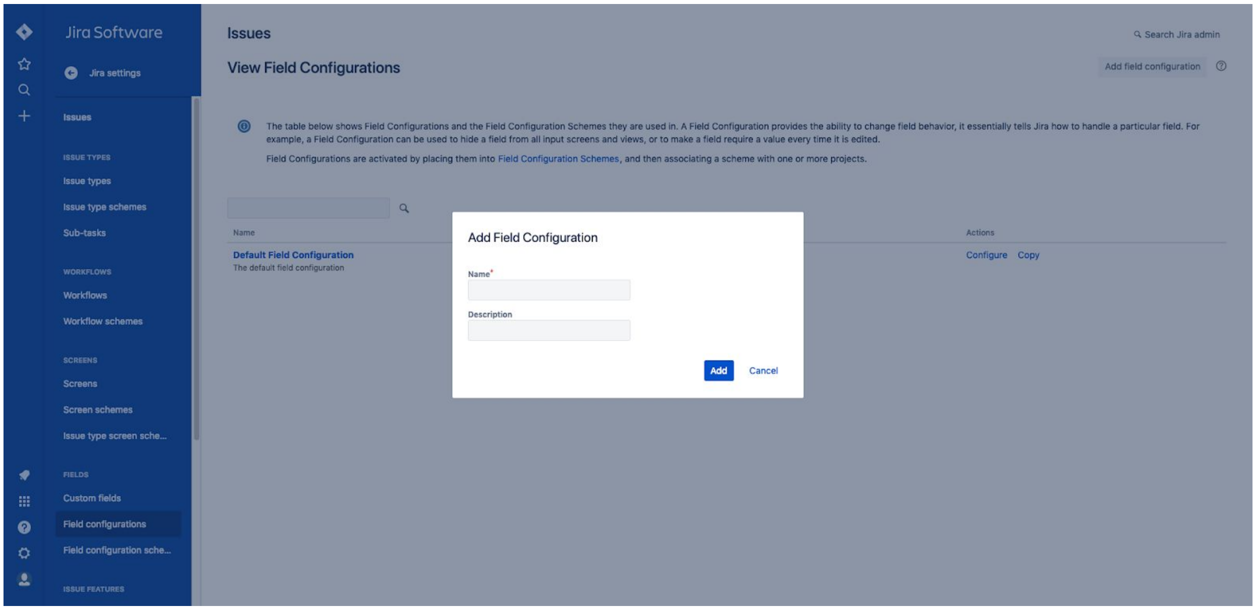
label × — **Labels**: Add arbitrary labels to issues.

13.34 — **Number Field**: A custom field that stores and validates numeric (floating point) input.

Next Cancel

Name		Available Context(s)	Screens
Approvers Contains users needed for appro		Issue type(s): Global (all issues)	
Change completion date Specify the completion time for t		Issue type(s): Global (all issues)	
Change reason Choose the reason for the chang		Issue type(s): Global (all issues)	
Change risk		Issue type(s): Global (all issues)	
Change start date Specify the time and date for the		Issue type(s): Global (all issues)	
Change type		Issue type(s): Global (all issues)	
Epic Color LOCKED Epic Color field for Jira Software		Issue type(s):	
Epic Link LOCKED Choose an epic to assign this iss		Issue type(s): Global (all issues)	DATA: Kanban Bug Screen DATA: Kanban Default Issue Screen
Epic Name LOCKED Provide a short name to identify t		Issue type(s):	DATA: Kanban Default Issue Screen
Epic Status LOCKED Epic Status field for Jira Software use only.	Status of Epic	Issue type(s):	
Impact	Select List (single choice)	Issue type(s): Global (all issues)	
Issue color LOCKED	Issue color	Issue type(s): Global (all issues)	

Figure 5. Creating a custom field.



Issues

View Field Configurations

The table below shows Field Configurations and the Field Configuration Schemes they are used in. A Field Configuration provides the ability to change field behavior, it essentially tells Jira how to handle a particular field. For example, a Field Configuration can be used to hide a field from all input screens and views, or to make a field require a value every time it is edited.

Field Configurations are activated by placing them into Field Configuration Schemes, and then associating a scheme with one or more projects.

Name	Actions
Default Field Configuration The default field configuration	Configure Copy

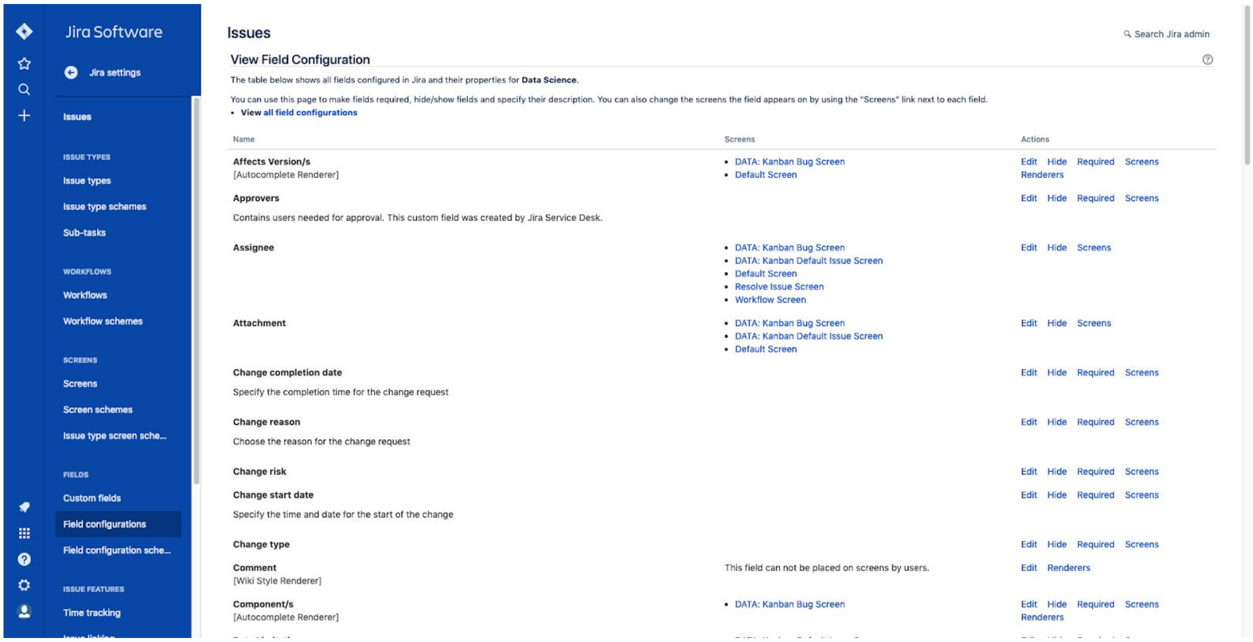
Add Field Configuration

Name*

Description

Add Cancel

Figure 6. Adding a field configuration.



Issues

View Field Configuration

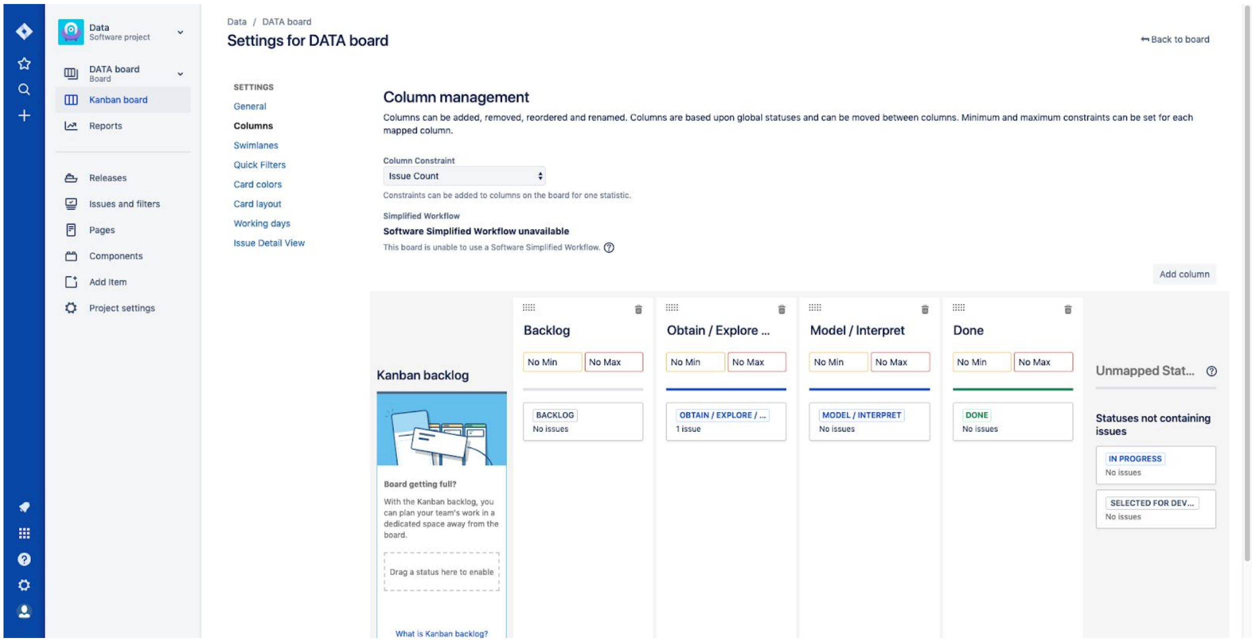
The table below shows all fields configured in Jira and their properties for **Data Science**.

You can use this page to make fields required, hide/show fields and specify their description. You can also change the screens the field appears on by using the "Screens" link next to each field.

- View **all field configurations**

Name	Screens	Actions
Affects Version/s [Autocomplete Renderer]	• DATA: Kanban Bug Screen • Default Screen	Edit Hide Required Screens Renderers
Approvers Contains users needed for approval. This custom field was created by Jira Service Desk.		Edit Hide Required Screens
Assignee	• DATA: Kanban Bug Screen • DATA: Kanban Default Issue Screen • Default Screen • Resolve Issue Screen • Workflow Screen	Edit Hide Screens
Attachment	• DATA: Kanban Bug Screen • DATA: Kanban Default Issue Screen • Default Screen	Edit Hide Screens
Change completion date Specify the completion time for the change request		Edit Hide Required Screens
Change reason Choose the reason for the change request		Edit Hide Required Screens
Change risk		Edit Hide Required Screens
Change start date Specify the time and date for the start of the change		Edit Hide Required Screens
Change type		Edit Hide Required Screens
Comment [Wiki Style Renderer]	This field can not be placed on screens by users.	Edit Renderers
Component/s [Autocomplete Renderer]	• DATA: Kanban Bug Screen	Edit Hide Required Screens Renderers

Figure 7. Editing a field configuration.



Data / DATA board

Settings for DATA board

SETTINGS
General
Columns
Swimlanes
Quick Filters
Card colors
Card layout
Working days
Issue Detail View

Column management

Columns can be added, removed, reordered and renamed. Columns are based upon global statuses and can be moved between columns. Minimum and maximum constraints can be set for each mapped column.

Column Constraint: Issue Count

Constraints can be added to columns on the board for one statistic.

Simplified Workflow

Software Simplified Workflow unavailable

This board is unable to use a Software Simplified Workflow.

Kanban backlog	Backlog	Obtain / Explore ...	Model / Interpret	Done	Unmapped Stat...
	No Min / No Max	No Min / No Max	No Min / No Max	No Min / No Max	Statuses not containing issues
Board getting full? With the Kanban backlog, you can plan your team's work in a dedicated space away from the board. Drag a status here to enable What is Kanban backlog?	BACKLOG No issues	OBTAIN / EXPLORE / ... 1 issue	MODEL / INTERPRET No issues	DONE No issues	IN PROGRESS No issues SELECTED FOR DEV... No issues

Figure 8. Changing the Data project's board to represent the Data Science workflow.