

# Методологии Data Science: текущие проблемы и будущие подходы

Иньиго Мартинеса, 6,\* Элизабет Вайлсдо нашей эры, Игорь Г Олайзолаа

*\*Фонд Vicomtech, Баскский исследовательский и технологический альянс (BRTA), Доностия-Сан-Себастьян 20009, Испания*

*бШкола инженерии ТЕКНУН, Университет Наварры, Доностия-Сан-Себастьян 20018, Испания*

*вИнститут науки о данных и искусственного интеллекта, Университет Наварры, Памплона 31009, Испания*

arXiv: 2106.07287v1 [cs.LG] 14 июня 2021 г.

## Абстрактный

Наука о данных приложила огромные усилия для разработки расширенной аналитики, улучшения моделей данных и создания новых алгоритмов. Однако не многие авторы сталкиваются с организационными и социально-техническими проблемами, которые возникают при выполнении проекта по науке о данных: отсутствие видения и четких целей, предвзятый акцент на технических вопросах, низкий уровень зрелости специальных проектов и неоднозначность ролей в науке о данных - одна из этих проблем. В литературе было предложено несколько методологий, посвященных этим типам проблем, некоторые из них относятся к середине 1990-х годов, и, следовательно, они не обновлены в соответствии с текущей парадигмой и последними разработками в области больших данных и технологий машинного обучения. Кроме того, меньшее количество методологий предлагает полное руководство для всей команды, проекта и управления данными и информацией. В этой статье мы хотели бы изучить необходимость разработки более целостного подхода к реализации проектов в области науки о данных. Сначала мы рассмотрим методологии, представленные в литературе, для работы над проектами в области науки о данных и классифицируем их в соответствии с их направленностью: проект, команда, управление данными и информацией. Наконец, мы предлагаем концептуальную основу, содержащую общие характеристики, которыми должна обладать методология управления проектами в области науки о данных с целостной точки зрения. Эта структура может использоваться другими исследователями в качестве дорожной карты для разработки новых методологий науки о данных или обновления существующих. Сначала мы рассмотрим методологии, представленные в литературе, для работы над проектами в области науки о данных и классифицируем их в соответствии с их направленностью: проект, команда, управление данными и информацией. Наконец, мы предлагаем концептуальную основу, содержащую общие характеристики, которыми должна обладать методология управления проектами в области науки о данных с целостной точки зрения. Эта структура может использоваться другими исследователями в качестве дорожной карты для разработки новых методологий науки о данных или обновления существующих. Сначала мы рассмотрим методологии, представленные в литературе, для работы над проектами в области науки о данных и классифицируем их в соответствии с их направленностью: проект, команда, управление данными и информацией. Наконец, мы предлагаем концептуальную основу, содержащую общие характеристики, которыми должна обладать методология управления проектами в области науки о данных с целостной точки зрения. Эта структура может использоваться другими исследователями в качестве дорожной карты для разработки новых методологий науки о данных или обновления существующих.

**Ключевые слова:** наука о данных, машинное обучение, большие данные, методология науки о данных, жизненный цикл проекта, влияние на организацию, управление знаниями, вычислительная среда

## 1. Введение

В последние годы область науки о данных привлекла к себе повышенное внимание и приложила большие усилия для разработки расширенной аналитики, улучшения моделей данных и развития новых алгоритмов. Последние достижения в этой области - хорошее отражение такого начинания [1]. Фактически, сообщество исследователей в области науки о данных растет день ото дня, исследуя новые области, создавая новые специализированные и экспертные роли и давая ростки все больше и больше ветвей процветающего дерева, называемого наукой о данных. Кроме того, дерево науки о данных не одиноко, так как оно питается соседними областями математики, статистики и информатики.

Однако эти недавние технические достижения не идут рука об руку с их применением в реальных проектах по науке о данных. В 2019 году VentureBeat [2] показал, что 87% проектов в области науки о данных никогда не попадают в производство, а исследование NewVantage [3] показало, что для 77% предприятий внедрение инициатив в области больших данных и искусственного интеллекта (ИИ) по-прежнему представляет собой серьезную проблему. Большой вызов. Также

[4] сообщили, что 80% аналитических данных не принесут результатов для бизнеса до 2022 года, а 80% проектов в области науки о данных «останутся алхимией, управляемой волшебниками» до 2020 года. Благодаря конкурентному преимуществу, которое такие передовые методы предоставляют исследователям и практикам, Примечательно видеть такие низкие показатели успешности проектов в области науки о данных.

Относительно того, как группы по анализу данных подходят и разрабатывают проекты в различных областях, Лео Брейман [5] утверждает, что существует две культуры статистического моделирования: а) ветвь прогнозирования, которая ориентирована на создание эффективных алгоритмов для получения хороших прогностических моделей для прогнозирования будущего. и б) ветвь моделирования, которая больше заинтересована в понимании реального мира и лежащих в основе процессов. С последней точки зрения теория, опыт, знание предметной области и причинно-следственная связь действительно имеют значение, и важен научный метод. С этой точки зрения философы науки XX века, такие как Карл Поппер, Томан Кун, Имре Лакатос или Пол Фейербенд, теоретизировали, как разворачивается процесс накопления знаний и науки [6]. Среди прочего, эти авторы обсуждали природу и происхождение научных идей, формулировка и использование научного метода, а также последствия различных методов и моделей науки. Несмотря на очевидные различия, в этой статье мы бы

\*Автор, ответственный за переписку

Адрес электронной почты: imartinez@vicomtech.org (Иньиго Мартинес), eviles@tecnun.es (Элизабет Вайлс), iolalizola@vicomtech.org (Игорь Г Олайзола)

хотели бы побудить специалистов по обработке данных обсудить формулировку и использование научного метода для исследовательской деятельности в области науки о данных, а также последствия различных методов для выполнения отраслевых и бизнес-проектов. В настоящее время наука о данных - это молодая область, производящая впечатление ручной работы. Однако, несмотря на всю неопределенность и исследовательский характер науки о данных, действительно может существовать строгая «наука», как ее понимали философы науки, и которой можно эффективно управлять и управлять ею.

В литературе несколько авторов [7, 8] сталкивались с организационными и социотехническими проблемами, которые возникают при выполнении проекта по науке о данных, например: отсутствие видения, стратегии и четких целей, предвзятый акцент на технических вопросах, отсутствие воспроизводимости и неоднозначности ролей среди этих проблем, которые приводят к низкому уровню зрелости проектов в области науки о данных, которые управляются в специальной манере.

Несмотря на то, что эти проблемы действительно существуют в реальных проектах по науке о данных, сообщество не слишком обеспокоено ими, и о решениях этих проблем написано недостаточно. Как будет показано в разделе 4, некоторые авторы предложили методологии для управления проектами в области науки о данных и придумали новые инструменты и процессы для решения указанных проблем.

Однако, несмотря на то, что предлагаемые решения являются лидером в решении этих проблем, реальность такова, что проекты по науке о данных не используют преимущества таких методологий. В опросе [9], проведенном в 2018 году среди профессионалов как из отрасли, так и из некоммерческих организаций, 82% респондентов не следовали явной методологии процесса разработки проектов в области науки о данных, но 85% респондентов считали, что использование улучшенный и более последовательный процесс позволил бы создать более эффективные проекты в области науки о данных.

Учитывая опрос, проведенный KDnuggets в 2014 г. [10], 43% респондентов использовали в качестве основной методологии CRISP-DM. Эта методология неизменно чаще всего используется в проектах по аналитике, интеллектуальному анализу данных и науке о данных, для каждого опроса KDnuggets, начиная с 2002 года и заканчивая самым последним опросом 2014 года [11]. Несмотря на свою популярность, CRISP-DM был создан еще в середине 1990-х и с момента своего создания не пересматривался.

Поэтому цель данной статьи - провести критический обзор методологий, которые помогают в управлении проектами в области науки о данных, классифицировать их в соответствии с их направленностью и оценить их компетенции в решении существующих проблем. В результате этого исследования мы предлагаем концептуальную основу, содержащую функции, которые могла бы иметь методология управления проектами в области науки о данных с целостной точки зрения. Эта схема может использоваться другими исследователями в качестве дорожной карты для расширения используемых в настоящее время методологий или для разработки новых.

С этого момента документ структурирован следующим образом: контекстуализация проблемы представлена в разделе 2, где даны некоторые определения терминов «наука о данных» и «большие данные», чтобы избежать всякого смысла.

устраните недоразумение и дополнительно опишите, о чем идет речь в проекте по науке о данных. В разделе 2 также представлены организационные и социально-технические проблемы, возникающие при выполнении проекта по науке о данных. Раздел 3 описывает методологию исследования, использованную в статье, и вводит вопросы исследования. В разделе 4 представлен критический обзор методологий проектов в области науки о данных. В разделе 5 обсуждаются полученные результаты и, наконец, в разделе 6 описываются направления дальнейшего расширения этого исследования и основные выводы статьи.

## 2. Теоретические основы

### 2.1. Предпосылки и определения

Чтобы избежать семантических недоразумений, мы начнем с более точного определения термина «наука о данных». Его определение поможет объяснить особенности проектов в области науки о данных и сформулировать методологию управления, предложенную в этой статье.

#### 2.1.1. Наука о данных

Среди авторов, которые объединяют науку о данных из давно сформировавшихся областей науки, существует общее согласие в отношении областей, которые подпитывают и развивают дерево науки о данных. Например, [12] определяет науку о данных как пересечение информатики, бизнес-инженерии, статистики, интеллектуального анализа данных, машинного обучения, исследования операций, шести сигм, автоматизации и предметной экспертизы, тогда как [13] утверждает, что наука о данных является междисциплинарным пересечением математических знаний, деловой хватки и хакерских навыков. Для [14] наука о данных требует навыков, варьирующихся от традиционной информатики до математики и искусства, и [15] представляет диаграмму Венна с наукой о данных, визуализированной как соединение а) хакерских навыков, б) знаний математики и статистики и в) существенного опыта. Напротив, для авторов [16], многие проблемы науки о данных являются проблемами статистической инженерии, но с более крупными и сложными данными, которые могут потребовать распределенных вычислений и методов машинного обучения в дополнение к статистическому моделированию.

Тем не менее, немногие авторы обращают внимание на фундаментальную цель науки о данных, которая имеет решающее значение для понимания роли науки о данных в областях бизнеса и промышленности и ее возможных доменных приложений. По мнению авторов [9], наука о данных - это анализ данных для решения проблем и разработки идей, в то время как [17] утверждает, что наука о данных использует «статистические методы и методы машинного обучения для больших многоструктурных данных в распределенной вычислительной среде для выявления корреляции и причинно-следственные связи, классифицируют и прогнозируют события, выявляют закономерности и аномалии и делают выводы о вероятностях, интересах и настроениях». Для них наука о данных объединяет опыт разработки программного обеспечения, управления данными и статистики. Как и в [18], наука о данных описывается как область, изучающая принципы вычислений,

Также растет интерес к определению работы, выполняемой специалистами по данным, и к перечислению необходимых навыков, чтобы стать специалистом по данным, чтобы лучше понять его роль среди традиционных должностей. В этом смысле [19] определяет специалиста по обработке данных как человека, который «лучше разбирается в статистике, чем любой инженер-программист, и лучше в разработке программного обеспечения, чем любой статистик», поскольку [20] представляет точку зрения «единорога» исследователей данных и утверждает, что данные ученые занимаются всем: от поиска данных, их обработки в масштабе, визуализации и написания рассказов.

Среди сотен различных интерпретаций, которые можно найти для науки о данных, мы берем комментарии, приведенные выше, в качестве справки и даем собственное определение, чтобы сформировать остальную часть статьи:

*Наука о данных - это междисциплинарная область, которая находится между информатикой, математикой и статистикой и включает использование научных методов и приемов для извлечения знаний и ценности из больших объемов структурированных и / или неструктурированных данных.*

Следовательно, из этого определения мы делаем вывод, что проекты по науке о данных нацелены на решение сложных реальных проблем с помощью методов, основанных на данных. В этом смысле наука о данных может применяться практически ко всем существующим секторам и областям: банковское дело (обнаружение мошенничества [21], моделирование кредитного риска [22], пожизненная ценность клиента [23]), финансы (сегментация клиентов [24], анализ рисков [25], алгоритмическая торговля [26]), здравоохранение (анализ медицинских изображений [27], открытие лекарств [28], биоинформатика [29]), оптимизация производства (прогнозирование отказов [30], планирование технического обслуживания [31], обнаружение аномалий [32]), электронная коммерция (таргетированная реклама [33], рекомендации по продукту [34], анализ настроений [35]), транспорт (беспилотные автомобили [36], управление цепочкой поставок [37], контроль перегрузки [38]), чтобы упомянуть некоторые из них.

#### 2.1.2. Технологии больших данных

Оглядываясь назад на эволюцию науки о данных в течение последнего десятилетия, ее быстрое развитие тесно связано с растущей способностью собирать, хранить и анализировать данные, генерируемые с возрастающей частотой [39]. Фактически, в середине 2000-х годов на каждом из этих этапов (сбор, хранение и анализ) произошли фундаментальные изменения, которые изменили парадигму науки о данных и больших данных.

Что касается сбора, рост доступных и надежных взаимосвязанных датчиков, встроенных в смартфоны и промышленное оборудование, значительно изменил подход к статистическому анализу. Фактически, традиционно стоимость приобретения была настолько высокой, что статистики тщательно собирали данные, чтобы быть необходимыми и достаточными для ответа на конкретный вопрос. Этот серьезный сдвиг в сборе данных действительно спровоцировал взрывной рост объема машинно-генерируемых данных. Что касается хранения, мы должны выделить разработку новых методов распространения

данные между узлами в кластере и разработка распределенных вычислений для параллельной работы с данными на этих узлах кластера. Экосистемы Hadoop и Apache Spark являются хорошими примерами новых технологий, которые внесли свой вклад в развитие методов сбора и хранения. Кроме того, важным прорывом в появлении новых алгоритмов и методов анализа данных стало увеличение вычислительной мощности как центральных, так и графических процессоров. В частности, последние достижения в области графических процессоров подтолкнули к расширению методов глубокого обучения, очень стремящихся к быстрым матричным операциям [40].

Помимо улучшений в области сбора, хранения и анализа, наука о данных получила пользу от огромного сообщества разработчиков и исследователей из передовых компаний и исследовательских центров. В нашей терминологии «большие данные», которые обычно определяют с помощью 5 V (объем, скорость, разнообразие, достоверность, ценность), - это часть области науки о данных, которая фокусируется на распределении данных и параллельной обработке.

В целом, в последние годы сообщество специалистов по науке о данных стремилось к совершенству и приложило огромные исследовательские усилия для разработки расширенной аналитики, сосредоточения внимания на решении технических проблем и, как следствие, организационных и социотехнических проблем. В следующем разделе обобщены основные проблемы, с которыми сталкиваются профессионалы в области науки о данных во время реальных деловых и отраслевых проектов.

## 2.2. Текущие проблемы

Использование науки о данных в контексте бизнес-организации сопряжено с дополнительными проблемами, помимо аналитических. Исследования, упомянутые во введении к этой статье, просто отражают существующие трудности в реализации проектов в области науки о данных и больших данных. Ниже мы собрали некоторые из основных проблем и проблем, которые возникают во время проекта по науке о данных, как на организационном, так и на техническом уровне.

*Координация, сотрудничество и общение* Область науки о данных эволюционирует от работы, выполняемой отдельными учеными, занимающимися данными «одиноким волком», к работе, выполняемой командой со специализированными дисциплинами. Рассматривая проекты по науке о данных как сложную командную работу, [41, 9, 42] поднимают *координация*, определяемого как «управление зависимостями между действиями задач», как самая большая проблема для проектов в области науки о данных. Плохо скординированные процессы приводят к путанице, неэффективности и ошибкам. Более того, отсутствие эффективной координации происходит как внутри групп анализа данных, так и во всей организации [43].

Помимо несогласованности, для [44, 45, 46] есть четкие *вопросы сотрудничества* и [47, 48] выделяют *отсутствие прозрачного общения* между тремя основными заинтересованными сторонами: бизнесом (клиентом), командой аналитиков и ИТ-отделом. Например, [44] упоминает о сложности развертывания аналитических команд в производственной среде, *координировать* с

*ИТ-отдел* и объяснить бизнес-партнерам науку о данных. [44] также свидетельствует об отсутствии поддержки со стороны бизнеса в том смысле, что для достижения хороших результатов не хватает бизнес-информации или экспертной информации в предметной области. В целом, похоже, что команда аналитиков данных и специалисты по данным в целом пытаются эффективно работать вместе с ИТ-отделом и бизнес-агентами.

Кроме того, [48] указывает *неэффективные модели управления* для анализа данных и [43] подчеркивают неадекватное управление и отсутствие спонсорской поддержки со стороны высшего руководства. В этом контексте [49] подтверждают, что работа в запутанной, хаотической среде может расстраивать и может снизить мотивацию членов команды и их способность сосредоточиться на целях проекта.

#### *Создание команд аналитики данных*

Другими словами, [50] выявляет проблемы, чтобы привлечь подходящую команду для проекта, а [45, 46, 43, 48, 51] выделяют *отсутствие людей с аналитическими навыками*. Эта нехватка специализированной аналитической рабочей силы заставила каждый крупный университет запускать новые программы по большим данным, аналитике или науке о данных [42]. В этой связи [46] выступает *запотребность в многопрофильной команде*: наука о данных, технологии, бизнес и управленческие навыки необходимы для достижения успеха в проектах по науке о данных. Например, в [9] говорится, что команды по обработке и анализу данных сильно зависят от ведущего специалиста по обработке данных, что связано с незрелостью процессов и отсутствием надежной командной методологии.

#### *Определение проекта по науке о данных*

Проекты в области науки о данных часто имеют весьма неопределенные входные данные, а также крайне неопределенные результаты и часто носят временный характер [52], предполагая значительный обмен мнениями между членами команды и методом проб и ошибок для определения правильных инструментов анализа, программы и параметры.

Исследовательский характер таких проектов делает *сложно установить адекватные ожидания* [17], установите реалистичные сроки проекта и оцените, сколько времени потребуется для завершения проекта [8]. В связи с этим [50, 53] отмечают, что масштаб проекта может быть трудно узнать заранее, а понимание бизнес-целей также является сложной задачей.

Более подробно авторы [47, 43, 48] выделяют *отсутствие четких бизнес-целей*, недостаточная рентабельность инвестиций или бизнес-кейсы, а также ненадлежащий масштаб проекта. Для [54] существует *предвзятый акцент на технических вопросах*, что ограничивает возможности организаций раскрыть весь потенциал аналитики данных. Вместо того, чтобы сосредоточить внимание на бизнес-проблеме, специалисты по данным часто были одержимы достижением современных результатов в задачах сравнительного анализа, но поиск небольшого увеличения производительности на самом деле может сделать модели слишком сложными, чтобы быть полезными. Такой образ мышления удобен для соревнований по науке о данных, таких как Kaggle [55], но не для промышленности. Соревнования Kaggle на самом деле отлично подходят для обучения машинному обучению, но они могут сформировать неверные ожидания относительно того, что требовать в реальных условиях бизнеса [56].

#### *Заинтересованные стороны против аналитики*

Кроме того, чаще всего проектное предложение четко не определено [44] и есть *недостаточное участие со стороны бизнеса*, который может просто предоставить данные и лишь часть информации о предметной области, предполагая, что команда аналитиков данных сделает остальную часть «магии» самостоятельно. Высокие ожидания, вызванные машинным обучением и методами глубокого обучения, вызвали ошибочное мнение о том, что эти новые технологии могут достичь всего, что предлагает бизнес, с очень низкими затратами, а это очень далеко от реальности [57]. Отсутствие участия со стороны бизнеса также может быть вызвано отсутствием взаимопонимания между обеими сторонами: специалисты по обработке данных могут не понимать предметную область данных, а бизнес обычно не знаком с методами анализа данных. По факту,

Выявленные проблемы управления проектом могут быть результатом *низкий уровень зрелости процесса* [52] и отсутствие последовательных методов и процессов для подхода к теме науки о данных [45]. Более того, последствия такого низкого уровня внедрения процессов и методологий могут также привести к поставке «неправильных вещей» [41, 9, 59] и «расширению масштабов» [41, 9]. Фактически, отсутствие эффективных процессов взаимодействия с заинтересованными сторонами увеличивает риск того, что команды создадут что-то, что не удовлетворяет потребности заинтересованных сторон. Наиболее очевидное изображение такой проблемы - нулевое воздействие *инеиспользование результатов проекта* со стороны бизнеса или клиента [44].

#### *Вождение с данными*

Использование подхода, основанного на данных, определяет основную особенность проекта в области науки о данных. Данные находятся в эпицентре всего проекта. Тем не менее, это также порождает некоторые конкретные вопросы, обсуждаемые ниже. Далее мы собрали основные проблемы, возникающие при работе с данными, независимо от того, связаны ли они с инструментами, с самой технологией или с управлением информацией.

Неоднократная жалоба специалистов по анализу данных в реальных проектах по обработке и анализу данных заключается в том, что *качество данных*: независимо от того, труднодоступны ли данные [46] или они «грязные» и содержат проблемы, специалисты по данным обычно приходят к выводу, что данные не обладают достаточным потенциалом, чтобы подходить для алгоритмов машинного обучения. Понимание того, какие данные могут быть доступны [50], их репрезентативность для рассматриваемой проблемы [46] и их ограничения [53], имеет решающее значение для успеха проекта. Фактически, [59] утверждает, что отсутствие скординированной очистки данных или проверки качества может привести к ошибочным результатам. В связи с этим специалисты по данным обычно забывают о стадии валидации. Чтобы обеспечить надежную проверку предлагаемого решения в реальных промышленных и деловых условиях, данные и / или опыт в предметной области должны собираться с достаточным упреждением.

Также важно учитывать перспективу больших данных [60]. Увеличение объема данных и скорости усиливается

требования к вычислениям и, следовательно, *зависимость проекта от ИТ-ресурсов* [8]. Кроме того, масштаб данных увеличивает сложность технологии и необходимую архитектуру и инфраструктуру [48] и, как следствие, соответствующие затраты [43].

Другими словами, [46, 43, 60] также подчеркивают важность *безопасность и конфиденциальность данных*, и [43, 48] указывают на сложную зависимость от устаревших систем и проблемы интеграции данных.

Что касается ограничений алгоритмов машинного обучения, одна из наиболее часто встречающихся проблем заключается в том, что популярные методы глубокого обучения требуют большого количества соответствующих обучающих данных, а их надежность постоянно ставится под сомнение. [51] поднимает *чрезмерные затраты на обучение и переподготовку моделей*. Фактически, специалисты по обработке данных, как правило, используют в 4 раза больше данных, чем им необходимо, для обучения моделей машинного обучения, что является ресурсоемким и дорогостоящим. Кроме того, [51] указывает, что специалисты по обработке данных часто могут сосредоточиться на неверные показатели производительности модели, без каких-либо ссылок на общие бизнес-цели или компромиссы между различными целями.

### Делитесь идеями

[41, 9, 60] указывают на проблему медленного обмена информацией и данными между членами команды. Они заявляют, что эти некачественные процессы для хранения, извлечения и обмена данными и документами тратят время, поскольку людям нужно искать информацию, и увеличивают риск использования неправильной версии. В связи с этим [41, 9, 59] обнаруживают *отсутствие воспроизводимости* в проектах по науке о данных. Фактически, они призывают к действиям и разработке новых инструментов для решения проблемы отсутствия воспроизводимости, поскольку может оказаться «невозможным дальнейшее развитие прошлых проектов из-за непоследовательного сохранения соответствующих артефактов», таких как данные, пакеты, документация и промежуточные результаты.

Это может стать серьезной проблемой для долгосрочной устойчивости проектов в области науки о данных. В нескольких проектах прикладной науки о данных основным результатом проекта может быть не модель машинного обучения или прогнозируемая величина интереса, а нематериальный объект, такой как сам процесс проекта или полученные знания в ходе его разработки. Хотя достижение целей проекта важно, в некоторых случаях более важно знать, как проект действительно достиг этих целей, по какому пути он шел, и понимать, почему он предпринял именно эти шаги, а не другие. Полученные знания о маршруте проекта в области науки о данных имеют решающее значение для понимания результатов и прокладывают путь для будущих проектов. Вот почему *этот знаниями нужно управлять и сохранять* в хорошем состоянии, и для этого решающее значение имеет способность воспроизводить задачи и эксперименты в области науки о данных. [51] утверждает, что сохранение институциональных знаний является проблемой, поскольку специалистов по обработке данных и разработчиков не хватает, и они могут перейти на новую работу.

Чтобы решить эту проблему, [51] предлагает все задокументировать и создать подробный реестр для всех новых моделей машинного обучения, что позволит будущим сотрудникам быстро воспроизвести работу, проделанную их предшественниками. В связи с этим [53]

отметили, что обмен знаниями в командах по анализу данных и во всей организации является одним из ключевых факторов успеха проекта, и [43, 48] также добавили управление данными и информацией.

В отношении управления данными [51] также указал на проблему множества похожих, но несовместимых наборов данных, когда внутри компании может циркулировать множество версий одних и тех же наборов данных без возможности определить, какой из них является правильным.

### Резюме

Представленные проблемные моменты были разделены на три основные категории в зависимости от того, относятся ли они к а) команде или организации, б) к управлению проектом или в) к управлению данными и информацией. Эта таксономия предназначена для облегчения и лучшего понимания типов проблем, возникающих во время проекта по науке о данных. Кроме того, эта классификация будет идти рука об руку с обзором методологий науки о данных, который будет представлен позже в документе. В [60] предлагается альтернативная система классификации проблем с большими данными, которая определяется проблемами данных, процессов и управления. Для них проблемы с данными связаны с характеристиками самих данных, проблемы с процессами возникают при обработке данных, а проблемы управления связаны с конфиденциальностью, безопасностью, управлением и отсутствием навыков. Мы утверждаем, что предлагаемая таксономия проблем включает эту классификацию и имеет более широкий взгляд. В таблице 1 приведены основные проблемы, возникающие при выполнении реальных проектов в области науки о данных.

Некоторые из проблем, перечисленных в таблице 1, считаются симптомом или отражением более крупной проблемы, которая заключается в отсутствии последовательной методологии в проектах по науке о данных, как было указано в [7]. В этом смысле [9] предположил, что расширенная методология обработки и анализа данных может повысить вероятность успеха проектов в области науки о данных. В той же статье автор представил опрос, проведенный в 2018 году среди профессионалов как из отрасли, так и из некоммерческих организаций, в котором 82% респондентов заявили, что они не следуют явной методологии процесса разработки проектов в области науки о данных, но 85% респондентов считают, что использование улучшенного и более последовательного процесса позволит создать более эффективные проекты в области науки о данных.

Поэтому в этой статье мы хотели бы изучить следующие вопросы исследования:

- RQ1: Какие методологии можно найти в литературе для управления проектами в области науки о данных?
- RQ2: Готовы ли эти доступные методологии к решению текущих задач?

### 3. Методология исследования.

Чтобы исследовать современное состояние методологий науки о данных, в этой статье мы сделали критический обзор литературы. Сондерс и Рохон [61] определяют

| Управление командой   | Управление проектом   | Управление данными и информацией  |
|---|---|---|
| Плохая координация  | Низкий уровень зрелости процессов   | Отсутствие воспроизводимости  |
| Проблемы сотрудничества между командами   | Неопределенные бизнес-цели Установление   | Сохранение и накопление знаний  |
| Отсутствие прозрачной коммуникации  | адекватных ожиданий Трудно установить   | Низкое качество данных для машинного обучения Отсутствие проверок качества                  |
| Неэффективные модели управления Отсутствие людей с аналитическими навыками Положитесь не только на ведущего специалиста по данным Создавайте мультидисциплинарные команды | реалистичные сроки проекта Пристрастный акцент на технических вопросах Доставить не то Проект, который не используется бизнесом | Нет данных проверки Безопасность и конфиденциальность данных Инвестиции в ИТ-инфраструктуру |

Таблица 1: Основные проблемы проектов в области науки о данных

критический обзор литературы как «сочетание наших знаний и понимания того, что было написано, наших навыков оценки и суждения, а также нашей способности четко и логично структурировать их в письменной форме». Они также указывают на несколько ключевых атрибутов критического обзора литературы: а) в нем обсуждаются и оцениваются наиболее актуальные исследования, относящиеся к теме, б) в нем признаются наиболее важные и актуальные теории, концепции и экспертов в данной области, в) контекстуализируется и обосновывается цели и задачи, и г) выявляются пробелы в знаниях, которые не исследовались в предыдущей литературе.

Представленный критический обзор литературы был проведен на основе предыдущих концепций и путем сравнения литературы по управлению проектами в области науки о данных. Основная причина для проведения критического обзора, а не систематического обзора, заключалась в том, что информация об использовании методологий науки о данных разбросана по разным источникам, таким как научные журналы, книги, а также блоги, официальные документы и открытые онлайн-платформы для публикации. Информация, доступная из неофициальных источников, действительно была очень важна для понимания перспектив реальных проектов в области науки о данных. Для отбора статей для включения был установлен набор критериев отбора:

- Источник и базы данных, использованные для поиска: Web of Science, Mendeley, Google Scholar, Google.
- Период времени: с 2010 по 2020 (исключение для CRISP-DM, с 1995 года)
- Английский язык
- Тип документа: журнальная статья, статья конференции, официальный документ. Использовались как академические, так и практические статьи. Также были включены статьи для практикующих специалистов, поскольку они могут дать представление о методологиях анализа данных на отраслевом уровне.
- Содержание: все отобранные статьи имели отношение к методологиям проектов в области науки о данных либо напрямую (заголовок или текст), либо косвенно (исходя из содержания). Все включенные документы содержали выводы, относящиеся как минимум к одной из трех категорий методологий, проанализированных в этом обзоре (т. Е. Групповые, проектные и данные и информационные аспекты управления проектами в области науки о данных).

В итоге для рассмотрения было отобрано 19 исследований (с 1996 по 2019 год). В каждом исследовании было много

информации, и поэтому было решено, что лучший способ сравнить исследования - создать сравнительную таблицу. В этой первой таблице была предпринята попытка разделить ключевые элементы исследований на четыре пункта:

1. Сведения о статье (Автор / Журнал / Год)
2. Основные идеи и ключевые слова
3. Перспектива (команда, проект, данные и информация)
4. Выводы

После этого была создана вторая таблица для анализа того, как каждая методология отвечала требованиям выявленных проблем. Эта таблица разделена на три отдельные таблицы [2,3,4], по одной для каждой категории проблем. В первом столбце учтены проблемы, указанные в разделе 2 для управления командой, управления проектами и управления данными и информацией. Чтобы справиться с этими проблемами, методологии разрабатывают и реализуют различные процессы. Примеры таких процессов включены во второй столбец. На основе предложения каждой методологии каждому критерию присваивается балл. Эти оценки пытаются отразить усилия, приложенные авторами для решения каждой задачи. Взяв за основу [62], была использована следующая система punctuation:

- 0: критерии не выполнены
- 1: критерии практически не выполняются
- 2: критерии выполнены частично
- 3: критерии выполнены в значительной степени

Количественная информация, содержащаяся в таблицах [2,3,4], также была проиллюстрирована на треугольном графике (рисунок 1 справа). Каждая ось представляет категорию проблем.

- команда, проект, данные и информация - а значение в процентах (%) отражает, насколько хорошо методология решает эти проблемы. Более конкретно, процентное значение рассчитывается как отношение между суммой баллов и максимально возможной суммой в каждой категории. Например, в категории «Управление проектами» 7 задач, поэтому максимально возможный балл -  $7 \cdot 3 = 21$ . Методология, набравшая 14 баллов, будет иметь процентную оценку  $14/21 = 66\%$ .

Процентные баллы, включенные в треугольный график, также использовались для оценки целостности методологий, которая изменяет, насколько хорошо данная методология охватывает все три категории проблем. Рисунок 1 (слева) иллюстрирует

оценки целостности рассмотренных методологий. Целостность рассчитывается как отношение площади треугольника к площади треугольника с точными оценками 100% ~~три категории~~, в которых явно есть макс.  $\text{imum area}$  ( $3/4$ ). Вычислить площадь треугольника с использованием трех процентных оценок непросто, как это объясняется в Приложении A.

Таким образом, в таблицах [2,3,4] количественно оцениваются качественные характеристики каждой методологии и ее пригодность для решения задач, указанных в разделе 2, для управления командами, проектами, данными и информацией.

#### 4. Результаты: критический обзор методологий анализа данных.

Ниже мы рассмотрим набор методологий, которые предлагаются руководящие принципы для управления проектами в области науки о данных и их выполнения. Каждая методология подробно описана, чтобы проанализировать, насколько хорошо она отвечает требованиям представленных задач. Сюда входит описание структуры, функций, принципов, артефактов и рекомендаций. Каждая методология завершается кратким описанием ее основных сильных и слабых сторон. Мы предлагаем читателю проверить оценку каждой методологии после прочтения соответствующего критического обзора.

##### 4.1. CRISP-DM

Межотраслевой стандартный процесс интеллектуального анализа данных (CRISP-DM) [63] - это открытая стандартная модель процесса, разработанная SPSS и Teradata в 1996 году, которая описывает общие подходы, используемые экспертами по интеллектуальному анализу данных. Он представляет собой структурированный, четко определенный и тщательно документированный итеративный процесс. CRISP-DM разбивает жизненный цикл проекта интеллектуального анализа данных на шесть этапов: понимание бизнеса, понимание данных, подготовка данных, моделирование, оценка и развертывание.

На этапе понимания бизнеса основное внимание уделяется целям проекта и требованиям с точки зрения бизнеса, и эта информация преобразуется в проблему науки о данных. Таким образом, он пытается согласовать цели бизнеса и науки о данных, устанавливая адекватные ожидания и фокусируясь на достижении того, чего ожидает бизнес. Результатом этого этапа обычно является фиксированный план проекта, который вряд ли учитывает сложность установления реалистичных сроков проекта из-за исследовательского характера проектов в области науки о данных и их внутренней неопределенности.

Остальные этапы (подготовка данных, моделирование, оценка и развертывание) довольно просты, и читатель наверняка знаком с ними. Строгое соблюдение процесса CRISP-DM вынуждает менеджера проекта документировать проект и решения, принятые в процессе, таким образом сохраняя большую часть полученных знаний. Кроме того, этап подготовки данных включает тесты для оценки качества данных, в то время как этап оценки придает заметное значение валидации и оценке результатов проекта.

Тем не менее, одним из основных недостатков CRISP-DM является то, что он не объясняет, как команды должны организовываться для выполнения определенных процессов, и не решает ни одной из вышеупомянутых проблем управления командой. В этом смысле, по словам [64], CRISP-DM нуждается в лучшей интеграции с процессами управления, требует согласования с программным обеспечением и методологиями гибкой разработки, а вместо простых контрольных списков ему также необходимо руководство по методам для отдельных действий в рамках этапов.

В целом CRISP-DM обеспечивает согласованную основу для руководств и документации по опыту. Жизненный цикл науки о данных, представленный CRISP-DM, обычно используется в качестве эталона для других методологий, которые воспроизводят его с различными вариациями. Однако CRISP-DM был задуман в 1996 году и поэтому не обновлен в соответствии с текущей парадигмой и последними достижениями в технологиях обработки и анализа данных, особенно в отношении достижений больших данных. По словам ветерана отрасли Грегори Пятецки из KD-Nuggets: «CRISP-DM остается самой популярной методологией для проектов аналитики, интеллектуального анализа данных и науки о данных, с долей 43% в последнем опросе KDnuggets [10], но заменяет не поддерживаемый CRISP -ДМ давно назрела ».

<sup>4</sup>Последовательный и хорошо документированный итеративный процесс

<sup>5</sup>Не объясняет, как команды должны организовываться и не решает вопросы управления командой

##### 4.2. Microsoft TDSP

Microsoft Team Data Science Process (TDSP) от Microsoft [65] - это «гибкая, итеративная методология анализа данных, которая помогает улучшить командную совместную работу и обучение». Он очень хорошо документирован и предоставляет несколько инструментов и утилит, облегчающих его использование. К сожалению, TDSP очень зависит от служб и политик Microsoft, и это затрудняет более широкое использование. Фактически, мы считаем, что любая методология должна быть независимой от каких-либо инструментов или технологий. Как было определено в [66], методология - это общий подход, который направляет методы и действия в определенной области с определенным набором правил, методов и процессов и не полагается на определенные технологии или инструменты. Благодаря независимости от инструментов Microsoft, на которые опирается TDSP, эта методология обеспечивает некоторые интересные процессы как для проекта, так и для команды, а также для данных и данных.

Ключевые компоненты TDSP: 1) определение жизненного цикла науки о данных 2) стандартизованная структура проекта 3) инфраструктура и ресурсы и 4) инструменты и утилиты для выполнения проекта.

Жизненный цикл проекта TDSP похож на CRISP-DM и включает пять итеративных этапов: а) понимание бизнеса, б) сбор и понимание данных, с) моделирование, д) развертывание, и е) принятие клиентов. Это действительно итеративный и циклический процесс: например, результаты этапа «Сбор и понимание данных» могут быть возвращены на этап «Бизнес-понимание».

Таблица 2: Результаты методологии управления проектами

| Для решения проблем ...                      | Методология предусматривает ...  | [4,1] [4,2] [4,3] [4,4] [4,5] [4,6] [4,7] [4,8] [4,9] [4,10] [4,11] [4,12] [4,13] [4,14] [4,15] [4,16] [4,17] ] [4,18] [4,19] |
|--|--|---|
| Низкий уровень зрелости процесса             | Жизненный цикл науки о данных: задачи высокого уровня / руководство / схема управления проектами                 | 3 3 3 3 3 0 3 3 2 3 3 3 3 2 2 3 3 3   |
| Неопределенные бизнес-цели                   | схема, чтобы не задавать важные вопросы слишком поздно   | 2 2 2 2 1 0 1 3 3 2 3 3 1 1 3 2 0 3 2   |
| Установите адекватные ожидания               | придает важность фазе понимания бизнеса или отрасли  | 3 3 2 3 2 0 2 3 3 2 2 3 2 3 2 2 1 2 3   |
| Трудно установить реалистичные сроки проекта | процессы для контроля и отслеживания продолжительности конкретных шагов  | 0 1 0 0 3 2 0 0 0 3 0 2 2 0 0 0 0 0 0   |
| Пристрастный акцент на технических вопросах  | участвует в согласовании бизнес-требований и целей науки о данных  | 2 2 3 2 0 0 1 3 3 3 2 0 2 1 2 2 1 2 2   |
| Доставка «не того»                           | гарантирует, что результат проекта соответствует ожиданиям клиента / исследователя                               | 2 3 1 2 1 0 0 3 3 2 2 3 2 3 0 2 0 2 2   |
| Результаты, не используемые бизнесом         | методы оценки результатов, интеграция услуги или продукта в клиентскую среду и обеспечивает необходимое обучение | 2 3 1 2 0 0 0 3 2 2 2 3 2 3 2 0 0 0 3 3   |
| Всего  |  | 14 17 12 14 10 2 7 18 16 17 14 17 14 14 11 10 5 15 15   |
| % выше идеального (21)                       |  | 67 57 67 48 10 33 86 76 67 67 67 67 67 52 48 24 71 71   |

Таблица 3: Результаты методологии управления командой

| Для решения проблем ...                                | Методология предусматривает ...  | [4,1] [4,2] [4,3] [4,4] [4,5] [4,6] [4,7] [4,8] [4,9] [4,10] [4,11] [4,12] [4,13] [4,14] [4,15] [4,16] [4,17] ] [4,18] [4,19] |
|--|--|---|
| Плохая координация                                     | устанавливает роли / обязанности в проекте по науке о данных                                     | 0 3 3 3 3 3 2 1 0 2 0 1 0 2 2 1 3 0   |
| Проблемы сотрудничества между командами                | определяет четкие задачи и проблемы для каждого человека, а также способы взаимодействия         | 0 3 3 3 3 2 1 0 2 0 1 0 2 2 1 3 0   |
| Отсутствие прозрачного общения                         | описывает, как команды должны работать, чтобы эффективно общаться, координировать и сотрудничать | 0 2 0 2 2 3 3 3 0 0 1 2 0 2 0 0 0 0 3 0 0   |
| Незэффективные модели управления                       | описывает, как координировать свои действия с ИТ, и приближается к этапу развертывания           | 0 0 2 3 2 3 0 1 3 2 1 3 1 0 1 3 0 2 3   |
| Положитесь не только на ведущего специалиста по данным | разделяет обязанности, способствует обучению сотрудников   | 0 1 2 1 2 2 0 2 0 1 2 1 1 0 2 0 0 2 0 0 2 0   |
| Создавайте мультидисциплинарные команды                | способствует совместной работе междисциплинарных профилей  | 0 2 1 2 1 2 1 0 1 1 1 3 0 0 2 1 2 0 1   |
| Всего  |  | 0 11 11 14 13 16 8 5 5 5 10 7 5 0 7 6 9 10 4  |
| % выше идеального (18)                                 |  | 0 61 61 78 72 89 44 28 год 42 56 39 28 год 0 39 33 50 56 22   |

Таблица 4: Результаты методологии управления данными и информацией

| Для решения проблем ...                                  | Методология предусматривает ...  | [4,1] [4,2] [4,3] [4,4] [4,5] [4,6] [4,7] [4,8] [4,9] [4,10] [4,11] [4,12] [4,13] [4,14] [4,15] [4,16] [4,17] ] [4,18] [4,19] |
|--|--|---|
| Отсутствие воспроизводимости                             | предлагает настройку для обеспечения воспроизводимости и прослеживаемости                | 0 3 3 1 2 2 3 0 3 0 0 0 0 0 0 3 0 2   |
| Сохранение и накопление порождения и накопления знаний   | 2 2 3 3 2 3 0 3 3 0 3 1 0 1 0 0 1 0 3 0 2  |   |
| Низкое качество данных для алгоритмов машинного обучения | значения: данные, модели, эксперименты, идеи проектов, лучшие практики и подводные камни | 1 0 0 2 0 0 0 0 0 0 0 0 0 0 0 2 1 0 1 2 1 2   |
| Отсутствие проверок качества                             | учитывает ограничения методов машинного обучения   | 2 3 0 2 1 0 1 0 0 3 2 0 2 2 0 2 2 0 2 2 3 2   |
| Нет данных проверки                                      | тесты для проверки ограничений по качеству и потенциальному использованию данных         | 2 2 3 2 0 0 3 1 2 3 2 1 1 2 0 0 0 0 0 0 0 2   |
| Безопасность и конфиденциальность данных                 | робастная проверка предложенного решения и выдвигает гипотезу                            | 0 0 0 2 0 0 2 0 0 2 0 0 2 0 0 0 0 0 0 0 0 2   |
| Инвестиции в ИТ-инфраструктуру                           | обеспечен безопасностью и конфиденциальностью данных                                     | 0 1 2 0 1 0 0 3 3 3 2 3 2 2 2 2 0 0 0 0 0 2   |
| Всего  |  | 7 11 11 12 8 5 9 6 10 11 7 6 9 6 7 5 10 8 12  |
| % выше идеального (21)                                   |  | 33 52 52 57 38 24 43 год 29 48 52 33 29 43 год 29 33 24 48 38 57  |

Система оценки: критерии не выполнены (0), с трудом выполнены (1), выполнены частично (2), выполнены в значительной степени (3).

Условные обозначения: 4.1 CRISP-DM; 4.2 Microsoft TDSP; 4.3 Жизненный цикл Domino DS; 4.4 PAMCIS; 4.5 Жизненный цикл Agile Data Science; 4.6 MIDST; 4.7 Рабочие процессы разработки для специалистов по данным; 4.8 Идея, оценка и внедрение больших данных; 4.9 Полотно управления большими данными; 4.10 Agile Delivery Framework; 4.11 Систематические исследования больших данных; 4.12 Структура управления большими данными; 4.13 Data Science Edge; 4.14 Основополагающая методология науки о данных; 4.15 Канва аналитики; 4.16 AI Ops; 4.17 Рабочий процесс Data Science; 4.18 Жизненный цикл EMC Data Analytics; 4.19 На пути к инженерии интеллектуального анализа данных

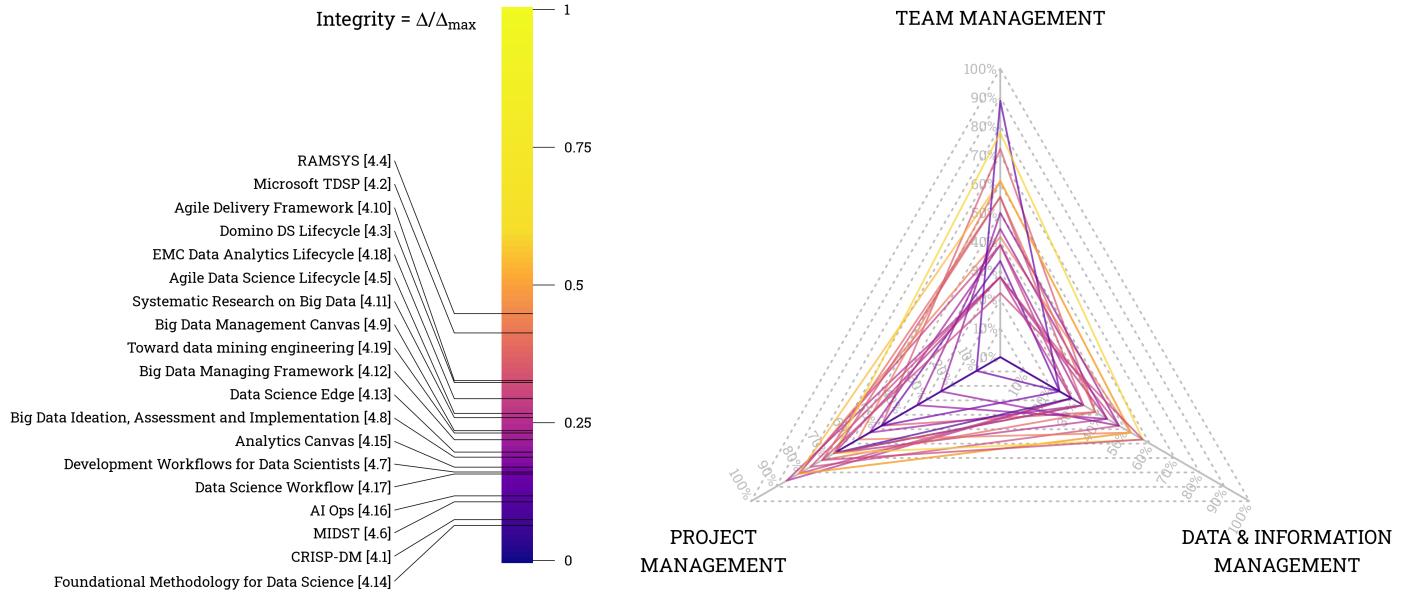


Рисунок 1: Количественное резюме рассмотренных методологий: а) значение целостности представлено на столбчатой диаграмме и б) оценки каждой категории показаны на треугольном графике, при этом цвет линии представляет целостность.

На этапе понимания бизнеса TDSP заботится об определении целей SMART (конкретные, измеримые, достижимые, релевантные, привязанные к срокам) и идентификации источников данных. В этом смысле одним из самых интересных артефактов является уставный документ: этот стандартный шаблон - это живой документ, который постоянно обновляется на протяжении всего проекта по мере того, как делаются новые открытия, а также меняются бизнес-требования. Этот артефакт помогает документировать процесс обнаружения проекта, а также способствует прозрачности и коммуникации, если в нем участвуют заинтересованные стороны. Кроме того, повторение уставного документа способствует получению и накоплению знаний и ценной информации для будущих проектов. Наряду с другими вспомогательными артефактами, этот документ может помочь проследить историю проекта и воспроизвести различные эксперименты. Что касается воспроизводимости, для каждой протестированной модели TDSP предоставляет отчет о модели: стандартный отчет на основе шаблона с подробной информацией о каждом эксперименте.

Что касается оценки качества, отчет о качестве данных готовится на этапе сбора и понимания данных. Этот отчет включает сводные данные, отношения между каждым атрибутом и целью, ранжирование переменных и многое другое. Чтобы упростить эту задачу, TDSP предоставляет автоматизированную утилиту IDEAR, которая помогает визуализировать данные и готовить сводные отчеты о данных. На заключительном этапе заказчик проверяет, соответствует ли система его бизнес-потребностям и отвечает ли она на вопросы с приемлемой точностью.

TDSP устраняет слабость CRISP-DM из-за отсутствия определения команды, определяя четыре отдельные роли (архитектор решения, менеджер проекта, специалист по данным и руководитель проекта) и их обязанности на каждом этапе проекта.

жизненный цикл. Эти роли очень хорошо определены с точки зрения управления проектами, и команда работает по методологиям Agile, которые улучшают сотрудничество и координацию. Их обязанности по созданию, реализации и развитию проекта ясны.

**4** Интегральная методология: обеспечивает процессы как на управление проектами, командой и данными и информацией

**5** Чрезмерная зависимость от инструментов и технологий Microsoft.

#### 4.3. Жизненный цикл Domino DS

Domino Data Lab представила свой жизненный цикл проекта в области науки о данных в официальном документе от 2017 года [59]. Вдохновленный CRISP-DM, Agile и серией наблюдений за клиентами, он использует «целостный подход ко всему жизненному циклу проекта от идеи до доставки и мониторинга». Методология основана на трех руководящих принципах: а) «Ожидайте и принимайте итерацию», но «не допускайте, чтобы итерации существенно задерживали проекты или отвлекали их от поставленной цели». б) «Разрешить сложное сотрудничество» путем создания компонентов, которые можно повторно использовать в других проектах. с) «Предвидеть потребности в возможности аудита» и «сохранять все соответствующие артефакты, связанные с разработкой и развертыванием модели».

Предлагаемый жизненный цикл науки о данных состоит из следующих этапов: а) создание идей; б) сбор и исследование данных; с) исследования и разработки; д) проверка; е) доставка; ф) мониторинг.

Идея отражает этап понимания бизнеса из CRISP-DM. На этом этапе определяются бизнес-цели, излагаются критерии успеха, и если это

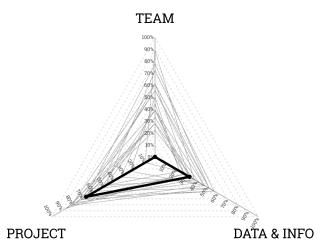


Рисунок 2: CRISP-DM

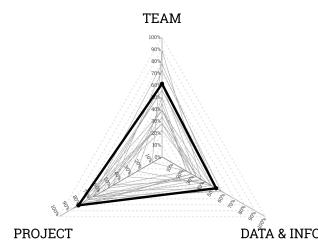


Рисунок 3: Microsoft TDSP

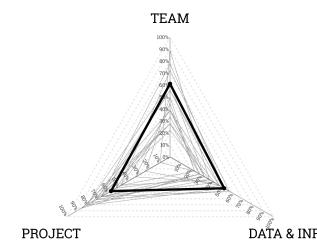


Рисунок 4: Жизненный цикл Domino DS

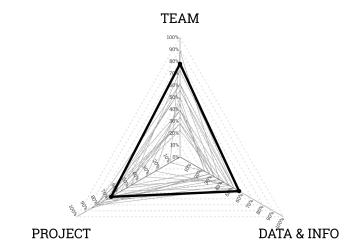


Рисунок 5: RAMSYS

В этом случае выполняется первоначальный анализ окупаемости инвестиций. На этом этапе также интегрируются общие методы гибкой разработки, включая разработку блога с учетом интересов заинтересованных сторон и создание готовых макетов. Эти предварительные действия по бизнес-анализу могут значительно снизить риск проекта за счет согласования всех заинтересованных сторон.

Этап сбора и исследования данных включает в себя многие элементы из этапов понимания и подготовки данных CRISP-DM. Этап исследований и разработок - это «воспринимаемое сердце» процесса обработки и анализа данных. Он повторяется через генерацию гипотез, экспериментирование и предоставление инсайтов. DominoLab рекомендует начинать с простых моделей, устанавливать частоту получения аналитических данных, отслеживать бизнес-ключевые показатели эффективности и устанавливать стандартные конфигурации оборудования и программного обеспечения, сохраняя при этом гибкость для экспериментов.

Этап проверки подчеркивает важность обеспечения воспроизводимости результатов, автоматических проверок и сохранения нулевых результатов. На этапе доставки Domino рекомендует сохранить связи между доставляемыми артефактами, пометить зависимости и разработать план мониторинга и обучения. Учитывая недетерминированный характер моделей, Domino предлагает методы мониторинга, которые выходят за рамки стандартной практики мониторинга программного обеспечения: например, использование контрольных групп в производственных моделях для отслеживания производительности модели и создания ценности для компаний.

В целом модель Domino не описывает каждый отдельный шаг, но более информативна, чтобы вести команду к лучшей производительности. Он эффективно объединяет науку о данных, разработку программного обеспечения и гибкие подходы. Более того, он использует гибкие стратегии, такие как быстрая итеративная доставка, тесное управление заинтересованными сторонами и отставание по продукту. По словам [67], жизненный цикл Domino не следует рассматривать как взаимоисключающий с CRISP-DM или Microsoft TDSP; скорее, его подход «передовой практики» с элементами «а ля карт» мог бы дополнить те или иные методологии, а не заменить их.

**4** Целостный подход к жизненному циклу проекта. Эффективно объединяет науку о данных, программную инженерию и гибкие подходы

**5** Информационная методология, а не предписывающая

#### 4.4. РАМСИС

RAMSYS [68] Стива Мойла - это методология поддержки проектов быстрого удаленного совместного анализа данных. Он предназначен для распределенных команд, и принципы, которыми руководствуются при разработке методологии, следующие: легкое управление, запуск в любое время, остановка в любое время, свобода решения проблем, обмен знаниями и безопасность.

Эта методология следует за методологией CRISP-DM и расширяет ее, а также позволяет прилагать усилия по интеллектуальному анализу данных в самых разных местах, обмениваясь данными через веб-инструмент. Цель методологии - обеспечить обмен информацией и знаниями, а также свободу экспериментировать с любой техникой решения проблем.

RAMSYS определяет три роли: «разработчики моделей», «мастер данных» и «комитет управления». «Мастер данных» отвечает за поддержку базы данных и применение необходимых преобразований. «Комитет управления» обеспечивает поток информации в сети и предоставление хорошего решения. Этот комитет также отвечает за управление интерфейсом с клиентом и постановку задачи, связанной с текущим проектом в области науки о данных: определение критериев успеха, получение и выбор представлений. Таким образом, «моделисты» экспериментируют, проверяют правильность каждой гипотезы и производят новые знания. При этом они могут предлагать новые преобразования данных «мастера данных». Одна из сильных сторон модели RAMSYS - это относительная свобода, предоставляемая моделям в проекте, чтобы пробовать свои собственные подходы.

«Информационное хранилище» - это предлагаемый артефакт, содержащий определение проблемы, данные и гипотезу о данных и моделях, а также другие концепции. Кроме того, РАМСИС предлагает использовать «инвестиционный счет гипотез», который содержит все необходимые операции для опровержения или подтверждения гипотезы: изложение гипотезы, свидетельство опровержения / подтверждения, доказательство, уточнение и обобщение гипотез. Этот артефакт можно использовать для извлечения многоразовых знаний в виде извлеченных уроков для будущих проектов.

Таким образом, RAMSYS намеревается дать возможность сотрудничать удаленно размещенным специалистам по данным в дисциплинированной манере, в том что касается потока информации, обеспечивая при этом свободный поток идей для решения проблем. Конечно, эту методологию можно применить к более общим данным.

научная группа, которая разделяет местоположение. Четкое определение ответственности для каждой роли обеспечивает эффективное сотрудничество, а инструменты, поддерживающие эту методологию, четко определены для применения в реальных проектах.

**4** Учитывает распределенные команды и предоставляет информацию и обмен знаниями

**5** Отсутствие поддержки обмена наборами данных и моделями

#### 4.5. Жизненный цикл Agile Data Science

Жизненный цикл Agile Data Science Lifecycle [69] Рассела Джерни представляет собой основу для выполнения науки о данных в сочетании с гибкой философией. Эта методология утверждает, что наиболее эффективный и подходящий способ сделать науку о данных ценным для организаций - это веб-приложение, и поэтому с этой точки зрения наука о данных превращается в создание приложений, которые описывают процесс прикладных исследований: быстрое прототипирование, исследовательский анализ данных, интерактивная визуализация и прикладное машинное обучение.

Книга включает в себя «Манифест Agile Data Science», в котором делается попытка применить гибкость в практике науки о данных и основан на нескольких принципах: 1) Итерировать, повторять, повторять: автор подчеркивает итеративный характер создания, тестирования и обучающие алгоритмы обучения. 2) Отправка промежуточных результатов: на протяжении циклов итерации любой промежуточный результат фиксируется и передается другим членам команды. В этом смысле, благодаря постоянному обмену работой, методология способствует обратной связи и созданию новых идей. 3) Эксперименты-прототипы над реализацией задач: неизвестно, приведет ли эксперимент к какому-либо ценному открытию. 4) Интегрируйте тиарическое мнение о данных в управление продуктом: важно прислушиваться к данным. 5) Поднимитесь вверх и вниз по пирамиде значений данных: пирамида значений данных обеспечивает путь процесса от начального сбора данных до обнаружения полезных действий. Создание ценности увеличивается по мере того, как команда поднимается на более высокие уровни пирамиды. 6) Найдите и следуйте критическому пути к продукту-убийце. Критический путь - это тот путь, который ведет к чему-то действенному, что создает ценность. 7) Опишите процесс, а не только конечное состояние. Задокументируйте процесс, чтобы понять, как был найден критический путь.

В целом Agile Data Science пытается согласовать науку о данных с остальной частью организации. Его основная цель - документировать и направлять исследовательский анализ данных, чтобы обнаружить и проследить критический путь к созданию привлекательного продукта. Методология также учитывает, что продукты создаются командами людей, и, следовательно, определяет широкий спектр командных ролей, от клиентов до инженеров DevOps.

**4** Быстрая доставка ценности от данных к клиенту.

Более реалистичная обратная связь: оцените ценность результатов

**5** Agile менее прост, лучше работает в динамике среды и меняющиеся требования

#### 4.6. MIDST

Кевин Кроустон [70] представляет теоретическую модель социотехнических возможностей для стигмергической координации. Цель этой методологии - улучшить координацию в командах по анализу данных путем передачи данных о координации от разработки бесплатного / бесплатного программного обеспечения с открытым исходным кодом (FLOSS). Для этого авторы разрабатывают и внедряют систему поддержки стигмергической координации.

Процесс является стигмергическим, если работа, проделанная одним агентом, создает стимул (стигму), который привлекает других агентов к продолжению работы. Таким образом, стигмергическая координация - это форма координации, основанная на сигналах совместной работы. Организация коллективных действий возникает из взаимодействия людей и развивающейся среды, а не из общего плана или прямого взаимодействия.

Как утверждается в статье, конкретные инструменты, которые были бы полезны командам специалистов по анализу данных, могут отличаться от функций, важных для групп разработчиков FLOSS, что затрудняет настройку среды совместной работы в области науки о данных с использованием существующих инструментов. Вот почему авторы предлагают альтернативный подход, основанный на стигмергической координации, и разработали веб-приложение для анализа данных для его поддержки. В связи с этим разработана теория аффорданса для поддержки стигмергической координации, которая основана на следующих принципах: видимость работы, использование четких жанров рабочих продуктов и сочетаемость вкладов.

Предлагаемая структура предназначена для улучшения сотрудничества между специалистами по данным и не пытается охватить управление проектами или вопросы управления данными. Следует учитывать, что представленное веб-приложение было дополнительном протестировано на реальных кейсах со студентами. Хотя верно то, что некоторые из рассмотренных методологий объясняются вместе с примерами использования, в большинстве из них отсутствуют эмпирические и полевые эксперименты. Именно поэтому особое внимание уделяется усилиям, прилагаемым данной методологией в этом направлении.

**4** Улучшает сотрудничество между специалистами по данным за счет стигмергическая координация.

**5** Исключительно нацелен на улучшение координации команды

#### 4.7. Рабочие процессы разработки для специалистов по данным

Рабочие процессы разработки для специалистов по данным [53] от Github и O'Reilly Media собирают различные передовые практики и рабочие процессы для специалистов по данным. В документе исследуется, как несколько организаций, управляемых данными, улучшают рабочие процессы разработки для науки о данных.

Предлагаемый процесс обработки и анализа данных следует итеративной структуре: а) задать интересный вопрос б) изучить предыдущую работу с) получить данные д) изучить данные е) смоделировать данные ф) проверить г) задокументировать код и) развернуть в производственной среде и) сообщить Результаты

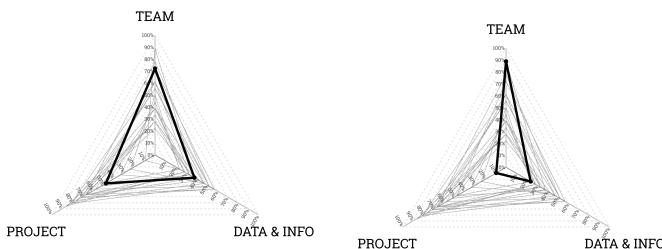


Рисунок 6: Жизненный цикл Agile Data Science

Рисунок 7: MIDST

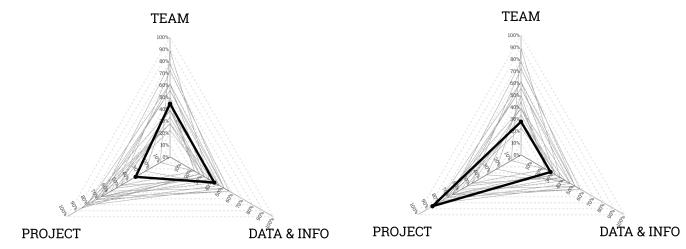


Рисунок 8: Рабочие процессы разработки для специалистов по данным

Рисунок 9: Идея, оценка и внедрение больших данных

Процесс начинается с того, что задается интересный вопрос, который описывается как одна из самых сложных задач в науке о данных. Прежде чем задавать интересные вопросы, необходимо понимать бизнес-цели и ограничения данных. В этом отношении определение подходящей меры успеха как для бизнеса, так и для команды специалистов по анализу данных также описывается как проблема. Следующим шагом в этом процессе является изучение предыдущей работы. Однако чаще всего команды по анализу данных сталкиваются с неструктурированными проектами и разрозненными знаниями, что затрудняет понимание предыдущей работы. В этом смысле они рекомендуют использовать инструменты, которые сделают работу науки о данных более доступной, например, базу знаний Airbnb [71].

С этого момента данные собираются и исследуются. Что касается сбора данных, авторы утверждают, что недостаточно просто собрать соответствующие данные, но также понять, как они были созданы, и решить вопросы безопасности, соответствия и анонимности. Чтобы помочь во время исследования данных, рекомендуется иметь хорошую структуру каталогов данных. Такие инструменты, как cookiecutter, могут позаботиться обо всех настройках и шаблонах для проектов в области науки о данных.

Что касается моделирования данных, предлагается создать две группы: одну для построения моделей и полностью независимую для оценки и проверки моделей. Таким образом предотвращаются утечки данных при обучении, а также обеспечивается защита критерии успеха от процесса создания модели. Следующий этап процесса - тестирование - пока еще обсуждается. Вот где расходятся практики науки о данных и разработки программного обеспечения: в науке о данных существует множество методов проб и ошибок, которые могут быть несовместимы с фреймворками, основанными на тестировании. Тем не менее, рекомендуется использовать технологии тестирования, чтобы улучшить интерпретируемость, точность и удобство использования.

Наличие фазы документации, без сомнения, является инновационным, но логичным моментом для методологий науки о данных. На самом деле документирования рабочего решения не всегда достаточно, так как не менее важно знать подводные камни и тупики. Рекомендуется создать кодовую книгу для записи всех предпринятых шагов, использованных инструментов, источников данных, результатов и сделанных выводов. Этот этап напрямую связан с этапом «изучения предыдущей работы», так как наличие хорошо задокументированного проекта поможет последующим проектам. Затем модели отправляются в производство. Для этой задачи предлагается использовать стандартный git-поток для версии всех изменений.

к моделям. Наконец, последний шаг в процессе - сообщить результаты, которые вместе с заданием исходного вопроса могут быть самыми проблемными шагами.

Эта методология построена на рекомендациях первоклассных организаций, управляемых данными, и в этом ее главная сила. Он включает в себя новые этапы рабочего процесса в области науки о данных, такие как «предыдущая проверка работы», «документация кода» и «передача результатов». Также предлагается структура команды и роли: специалист по данным, инженер по машинному обучению, инженер по данным. В целом можно сделать вывод, что в конечном итоге хороший рабочий процесс зависит от задач, целей и ценностей каждой команды, но он рекомендуется быстро получать результаты, воспроизводить, повторно использовать и проверять результаты, а также обеспечивать совместную работу и обмен знаниями.

**4 Основано на рекомендациях и передовом опыте первоклассные организации, управляемые данными**

**5 Отсутствует подробная разбивка каждой фазы**

#### 4.8. Идея, оценка и внедрение больших данных

Идея, оценка и реализация больших данных Мартином Ванауэром [57] - это методология, помогающая генерировать идеи больших данных, оценивать идеи и управлять их реализацией. Он основан на четырех аспектах больших данных (объем, разнообразие, скорость, ценность и достоверность), теории ценности ИТ, процессах формирования идей рабочих групп и управлении архитектурой предприятия.

Авторы утверждают, что внедрение больших данных напоминает инновационный процесс, и поэтому они следуют модели инноваций в рабочих группах, чтобы предложить свою методологию, которая разделена на две фазы: создание идеи и внедрение. Идея означает создание решений путем применения больших данных в новых ситуациях. Внедрение связано с оценкой разработанных решений и последующей реализацией.

Более конкретно, для фазы создания идей определены две перспективы: либо существуют бизнес-требования, которые могут быть лучше выполнены ИТ-специалистами, которые авторы назвали «Бизнес прежде всего» (BF), либо ИТ-отдел открывает новые возможности для бизнеса, так называемые «Данные превыше всего» (DF). С этой точки зрения не существует такой вещи, как «Технологии прежде всего». Авторы утверждают, что приоритизация технологий неосуществима, потому что технология должна быть представлена не как самоцель, а так, чтобы она приносила пользу.

Таким образом, этап формирования идеи напоминает либо определение потребностей бизнеса (BF), либо разработку новых бизнес-моделей на основе идентификации доступных данных (DF).

Каждая фаза (идея и реализация) также структурирована на подэтапы перехода и действия: 1) Переход идеи: включает определение миссии и идентификацию целей. Для поддержки этого этапа используются два артефакта: «Моделирование и разработка требований» для BF и «Оценка ключевых ресурсов» для перспективы DF. 2) Идея действий: здесь творческие решения определяются путем подготовки бизнес-сценария использования (BF) или ценностного предложения (DF) с помощью «холста бизнес-модели». 3) Переход к реализации: включает оценку и выбор наиболее подходящих идей. Здесь проводится финансовое и организационное технико-экономическое обоснование с помощью анализа затрат / выгод (BF) или оценки соответствия ценностного предложения (DF). Техническая осуществимость также анализируется с использованием методов EAM для оценки воздействия и развертывания. 4) Действие по реализации: наконец, разработана дорожная карта реализации.

Это предложение представляет собой совершенно новую перспективу, не имеющую ничего общего с остальными из представленных методологий. Он показывает совершенно другой конвейер, который отличается от жизненных циклов аналитики данных на основе CRISP-DM. Его основная сила заключается в структуре этапов разработки и реализации, но, что более важно, в различии перспектив бизнес-приоритета и приоритетов данных, которые определяют последующие пути, по которым следует идти. Однако он не решает ни задачи, связанные с командой, ни вопросы управления данными.

<sup>4</sup> Различие между перспективами, ориентированными на бизнес, и первоочередными данными

<sup>5</sup> Не учитывает вопросы, связанные с командой и управлением данными

#### 4.9. Канва управления большими данными

Canvas управления большими данными, созданный Майклом Кауфманном [72], представляет собой эталонную модель для управления большими данными, которая реализует создание ценности на основе данных путем связывания бизнес-целей с технической реализацией.

В этой исследовательской статье предлагается процесс управления большими данными, который помогает формировать знания и создавать ценности. Кауфманн утверждает, что создание ценности из больших данных зависит от «поддержки решений с помощью новых знаний, извлеченных из анализа данных». Он обеспечивает основу для науки о данных, чтобы ориентировать разработку информационных систем для обработки больших данных на создание знаний и ценностей. Автор включает термин ценности в парадигму больших данных, в данном случае определяемую 5V: объем, скорость, разнообразие, достоверность и ценность.

В этом отношении управление большими данными относится к процессу «управления потоками больших объемов, высокоскоростных, разнородных и / или неопределенных данных для создания ценности». Для обеспечения того, чтобы проект был эффективно сфокусирован на создании ценности, согласование бизнеса и ИТ является важным принципом. Представленное полотно управления большими данными состоит из

следующие пять этапов: 1) Подготовка данных: объединение данных из разных источников в единую платформу с постоянным доступом для аналитики 2) Аналитика данных: извлечение практических знаний непосредственно из данных посредством процесса обнаружения, формулирования гипотез и проверки гипотез. 3) Взаимодействие с данными: определение того, как пользователи влияют на результаты анализа данных. 4) Выполнение данных: использование результатов анализа данных для создания ценности в продуктах и услугах. 5) Интеллектуальные данные: способность организации приобретать и применять знания и навыки в управлении данными. В этом смысле существует три типа знаний и навыков: а) знания, полученные из данных; б) знания о данных и управлении данными; в) знания и навыки, необходимые для анализа и управления данными.

Эта структура основана на эпистемической модели управления большими данными как когнитивной системы: она предполагает, что знания возникают не в результате пассивного наблюдения, а в результате «итеративных замкнутых циклов преднамеренного создания и наблюдения изменений в окружающей среде». Ориентируясь на это видение аналитики данных, он утверждает, что знания, которые генерирует аналитика больших данных, возникают в результате взаимодействия специалистов по данным и конечных пользователей с существующими базами данных и результатами анализа.

В целом, эта методология предлагает иную точку зрения на науку о данных, явно отдавая приоритет созданию ценности на основе данных. Большинство подходов к исследованию больших данных нацелены на хранение, вычисления и аналитику данных, а не на знания и ценность, которые должны быть результатом обработки больших данных. Это изменение точки зрения влияет на развитие проекта, постоянно уравновешивая и согласовывая бизнес и технологии.

<sup>4</sup> Отдает приоритет созданию ценности на основе данных, а не на основе данных.  
советы по хранению, вычислениям и аналитике.

#### 5 Проблемы управления командой не решаются

#### 4.10. Фреймворк гибкой доставки

Ларсон и Чанг [73] предлагают структуру, основанную на синтезе гибких принципов с бизнес-аналитикой (BI), быстрой аналитикой и наукой о данных. Есть два уровня стратегических задач: (A) верхний уровень включает доставку бизнес-аналитики и (B) нижний уровень включает быструю аналитику и анализ данных.

На верхнем уровне есть пять последовательных шагов: обнаружение, проектирование, разработка, развертывание и предоставление ценности. A1) «обнаружение» - это когда заинтересованные стороны определяют бизнес-требования и определяют операционные границы. A2) «дизайн» фокусируется на моделировании и создании архитектуры системы. A3) «разработка» - это очень широкая фаза, которая включает в себя широкий спектр действий, например, кодирование ETL или создание сценариев для планирования заданий. A4) «развертывание» фокусируется на интеграции новых функций и возможностей в производственную среду. A5) «предоставление ценности» включает обслуживание, управление изменениями и обратную связь с конечными пользователями.

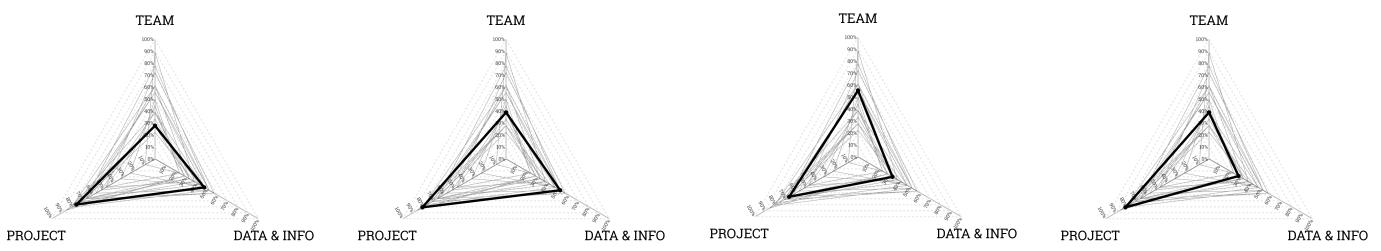


Рисунок 10: Область управления большими данными

Рисунок 11: Agile Delivery Framework

Рисунок 12: Систематические исследования больших данных

Рисунок 13: Структура управления большими данными

Нижний уровень включает шесть последовательных шагов: объем, сбор данных, анализ, разработка модели, проверка и развертывание. В1) «объем» определяет постановку проблемы и объем источников данных. В2) «сбор данных» собирает данные из озера данных и оценивает ценность источников данных. В3) «анализ» визуализирует данные и создает отчет о профилировании данных. В4) «разработка модели» подгоняет статистические модели и модели машинного обучения к данным. В5) «проверка» подтверждает качество модели. И разработка модели, и проверка выполняются с помощью итераций с ограничением по времени и гибких методов. В6) «развертывание» подготавливает информационные панели и другие инструменты визуализации.

Эта структура предназначена для поощрения успешного сотрудничества бизнеса и заинтересованных сторон в сфере ИТ. По мнению авторов, наука о данных по своей сути является гибкой, поскольку процесс выполняется с помощью итераций, а группы по анализу данных обычно состоят из небольших групп и требуют сотрудничества между бизнес-партнерами и техническими экспертами. Точно так же жизненный цикл бизнес-аналитики состоит из трех этапов, в которых может быть подходящим гибкий подход: этапы обнаружения, проектирования и разработки могут выиграть от итерационных циклов и небольших временных интервалов, даже при отсутствии каких-либо задач по программированию программного обеспечения. Основным новаторским подходом этой методологии является полное разделение мира бизнес-аналитики и анализа данных. По факту,

#### **4 Две методологии бизнес-аналитики и анализ данных, которые работают параллельно**

#### **5 Проблемы управления данными и воспроизводимость-иски опущены**

##### *4.11. Систематические исследования больших данных*

Das et al. [17] исследуют систематизацию исследовательских практик, основанных на данных. Представленный процесс состоит из восьми аналитических шагов Agile, которые начинаются с заданного набора данных и заканчиваются результатами исследования. В этом смысле [74] является рекомендуемым стилем разработки процесса.

Сначала информация извлекается и очищается. Несмотря на то, что в основе исследования, основанного на данных, лежат данные, авторы заявляют, что процесс должен начинаться с вопроса о том, что предполагается получить в результате исследования и данных.

Следовательно, следующий этап - это предварительный анализ данных, в котором из данных выявляются некоторые закономерности и информация. Эти шаблоны могут быть извлечены с помощью неконтролируемых методов и будут использоваться для поиска подходящей цели и отправной точки исследования. Следовательно, сгенерированные шаблоны данных и открытия приводят к определению цели исследования или исследовательской гипотезы.

После определения цели исследования можно будет продолжить работу с извлеченными данными. В противном случае цель исследования должна быть изменена, и задачи предварительного анализа данных, такие как описательная аналитика, должны быть выполнены снова. Таким образом, можно провести дальнейший анализ, учитывая цель исследования, выбрав наиболее подходящие функции и построив модели машинного обучения. Выходные данные этих моделей или систем прогнозирования подвергаются дальнейшей оценке. В рамках Agile-процесса оценка результатов может выполняться с помощью итеративных улучшений. Наконец, важно эффективно сообщать и сообщать о результатах, используя значимую инфографику и методы визуализации данных. Эти шаги могут повторяться итеративно до тех пор, пока не будут достигнуты желаемые результаты или уровень производительности.

Эта методология предлагает выполнить описанный выше рабочий процесс с небольшим подмножеством набора данных, выявить проблемы на ранней стадии и только в случае удовлетворительного результата расширить до всего набора данных. Авторы утверждают, что помимо использования итеративного гибкого планирования и выполнения, обобщенный набор данных и стандартизованная обработка данных могут сделать исследования, основанные на данных, более систематичными и последовательными.

В целом, эта структура обеспечивает процесс проведения систематических исследований больших данных и поддерживается гибкими методологиями в процессе разработки. Agile может способствовать частому сотрудничеству и развитию, основанному на ценностях, итеративным, поэтапным способом. Однако авторы не прилагают достаточно усилий для определения ролей и взаимодействия членов команды на разных этапах проекта. В целом, эта методология ближе к исследовательскому и академическому миру, и, поскольку наука о данных находится между исследовательской деятельностью и бизнес-приложениями, очень ценно иметь методологию, позволяющую извлекать лучшее из хотя бы одного из миров.

#### **4 Исследовательский и системный подход**

#### **5 Не определяет роли и взаимодействия между командой участников на разных этапах проекта**

#### 4.12. Платформа управления большими данными

Dutta et al. [75] представляют новую структуру для реализации проектов аналитики больших данных, которая сочетает в себе аспекты управления изменениями в структуре управления ИТ-проектами с аспектами управления данными в аналитической структуре. Эта методология представляет собой итеративный и циклический процесс для проекта больших данных, который разделен на три отдельных этапа: стратегическая основа, анализ данных и реализация.

Фаза стратегической подготовки включает определение бизнес-проблемы, мозговой штурм, концептуализацию решения, которое будет принято, и формирование проектных команд. Более подробно, этап «бизнес-проблемы» устанавливает правильные ожидания заинтересованных сторон и развеивает любые мифы о том, чего можно достичь в проекте больших данных. На этапе «исследования» изучается, как были решены похожие проблемы, а также ищутся различные аналитические продукты, доступные на рынке. Затем команда сформирована, и она рекомендует создавать ее с людьми из разных слоев общества: бизнес-подразделения, ИТ-эксперты, специалисты по обработке данных, эксперты в данной области, лица, принимающие бизнес-решения и т. д. Таким образом, эта структура выступает за кросс-функциональную и междисциплинарную команды. При этом готовится дорожная карта проекта, сбор основных мероприятий проекта с указанием сроков и назначенных людей. Эта дорожная карта должна быть достаточно гибкой и фокусироваться на выполнении и доставке, а не на строгом следовании плану.

Этап анализа данных имеет четкую структуру, состоящую из «сбора и изучения данных», «анализа и моделирования данных», «визуализации данных» и «генерации инсайтов». Как отмечается в других методологиях, широко используется предварительная фаза для сбора и исследования данных. Тем не менее, использование методов визуализации данных для представления результатов задач анализа данных является новым моментом и указывает на то, что авторы очень озабочены вовлечением всех заинтересованных сторон в разработку проекта. Кроме того, инновационное визуальное представление данных может помочь в генерации понимания и обнаружении схожих и аномальных закономерностей в данных. Этап анализа данных завершается этапом генерации информации, который является еще одним новаторским моментом, который иногда не учитывается другими методологиями. Важно понимать скрытые причины, лежащие в основе тенденций изменения данных. Этот последний шаг приводит анализ к пониманию и возможной практической бизнес-информации, которая может иметь ценность для организации.

Наконец, на этапе внедрения решение интегрируется с ИТ-системой, и необходимые люди проходят обучение, чтобы научиться его использовать. Интеграция - это этап, подверженный тренировке, поскольку он включает в себя работу с существующими ИТ-системами, а создание надежной архитектуры - сложная задача. Чтобы упростить внедрение новой системы, пользователи должны быть «обучены тому, как использовать инструменты и доступные данные». Таким образом, пользователи будут чувствовать себя более комфортно с новым источником информации при принятии бизнес-решений.

**4** Подчеркивает аспекты управления изменениями, такие как перекрестные формирования функциональной команды и обучение людей

**5** Недостаточно внимания уделяется валидации модели анализа данных и машинного обучения

#### 4.13. Граница науки о данных

Методология Data Science Edge (DSE) представлена в двух статьях Grady et al. [64, 76]. Это усовершенствованная модель процесса, подходящая для технологий больших данных и деятельности в области науки о данных. DSE обеспечивает полный жизненный цикл аналитики с пятиэтапной моделью - планирование, сбор, курирование, анализ, действие, которые организованы в зависимости от зрелости данных для удовлетворения требований проекта. Авторы делят жизненный цикл на четыре сектора: оценка, проектирование, строительство и улучшение, которые подробно описаны ниже.

Первый квадрант - оценка - состоит из планирования, анализа альтернатив и приблизительного порядка оценки процесса. Он включает требования и определение любых критических факторов успеха. Действия в рамках этого первого этапа оценки соответствуют этапу «планирования»: определение организационных границ, обоснование инвестиций и решение проблем, связанных с политикой и управлением.

«Архитектор» - это второй квадрант, который заключается в переводе требований в решение. В эту фазу входят этапы «сбор» и «курирование». В частности, этап «сбора» отвечает за управление базами данных, распределение данных и динамический поиск. Аналогичным образом, на этапе «курирования» собираются следующие действия: исследовательская визуализация, действия по обеспечению конфиденциальности и слияния данных, а также проверки оценки качества данных, среди прочего.

Квадрант «сборка» включает этапы «анализа» и «действия» и состоит из разработки, тестирования и развертывания технического решения. Новыми действиями на этапе являются: поиск более простых вопросов, задержка и параллелизм или проблема корреляции и причинно-следственной связи. Последний квадрант, «улучшение», состоит из эксплуатации и управления системой, а также анализа «инновационных способов повышения производительности системы».

В целом Grady et al. предложили новую модель процесса анализа больших данных, которая расширяет CRISP-DM и предоставляет полезные усовершенствования. Утверждается, что DSE служит полным жизненным циклом открытия знаний, который включает хранение, сбор данных и разработку программного обеспечения. Авторы объясняют, как модель процесса DSE согласуется с гибкими методологиями, и предлагают необходимые концептуальные изменения для внедрения гибкой разработки в аналитику данных. Наконец, авторы приходят к выводу, что, следуя гибким методологиям, можно ожидать более быстрых результатов для принятия решения и подтверждения текущего состояния проекта.

**4** Усовершенствованная модель процесса CRISP-DM с учетом датировать технологии больших данных и деятельность в области науки о данных

**5** Проблемы, связанные с управлением командой, опущены.

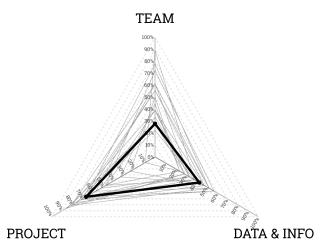


Рисунок 14: Граница науки о данных

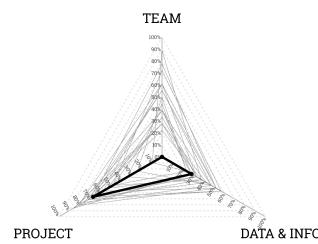


Рисунок 15: FMDS

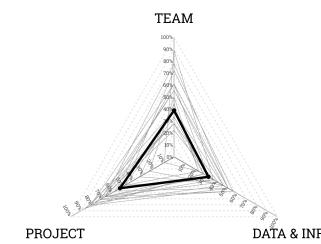


Рисунок 16: Область аналитики

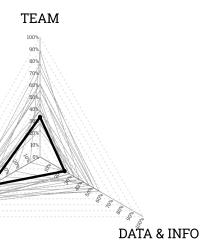


Рисунок 17: AI Ops

#### 4.14. Основополагающая методология науки о данных

Основополагающая методология науки о данных [77] от IBM имеет некоторое сходство с CRISP-DM, но предлагает ряд новых практик. Десять шагов FMDS иллюстрируют итеративный характер процесса обработки данных, в котором несколько этапов объединены замкнутыми циклами. Десять этапов говорят сами за себя: 1) понимание бизнеса 2) аналитический подход 3) требования к данным 4) сбор данных 5) понимание данных 6) подготовка данных 7) моделирование 8) оценка 9) развертывание 10) обратная связь.

Следует отметить интересное положение фазы аналитического подхода в начале проекта, сразу после понимания бизнеса и без сбора данных или исследовательского анализа. Несмотря на то, что точка зрения автора понятна, в действительности специалистам по данным трудно выбрать аналитический подход до исследования данных. Постановка бизнес-проблемы в контексте методов статистического и машинного обучения обычно требует предварительного исследования с использованием описательной статистики и методов визуализации. В целом эта методология структурирует проект по науке о данных в несколько этапов (10), чем CRISP-DM (6), но имеет те же недостатки, что и отсутствие определения различных ролей команды, а также отсутствие проблем с воспроизводимостью и накоплением знаний. и безопасность данных.

4 Предлагает новые практики и расширяет возможности CRISP-DM про- модель сбора

5 Наследует некоторые недостатки CRISP-DM, особенно полностью о команде и управлении данными

#### 4.15. Аналитика Canvas

Канва аналитики Арно Куна [47] - это полуформальная методика спецификации для концептуального проектирования проектов анализа данных. В нем описывается пример использования аналитики и документируется необходимая инфраструктура данных на раннем этапе планирования аналитического проекта. Он предлагает четырехуровневую модель, которая различает аналитический вариант использования, анализ данных, путь данных и источники данных.

Первым шагом является понимание предметной области и определение варианта использования аналитики: анализ первопричин, мониторинг процессов, профилактическое обслуживание или управление процессами. Этой стадии обычно инициируют агенты управления, так как они имеют широкое видение компании. На основе

В сценарии использования аналитики источники данных указываются экспертами в предметной области: датчики, системы управления, ERP-системы, CRM и т. д. Помимо знания того, откуда поступают данные, необходимо указать места, где данные хранятся. Для этого назначается ИТ-специалист. Наконец, вариант использования аналитики связан с соответствующей задачей анализа данных: описательной, диагностической, прогнозирующей или предписывающей. Здесь специалист по данным - главный агент.

Таким образом, аналитический холст назначает особую роль каждой фазе: руководство, эксперт в предметной области, ИТ-эксперт, специалист по данным. Команду также курирует архитектор аналитики. Этот холст очень полезен для структурирования проекта на ранних этапах и определения его основных конструкций. Включение такого инструмента помогает ставить четкие цели и способствует прозрачному общению между заинтересованными сторонами. Помимо описания сценария использования аналитики и необходимой инфраструктуры данных, панель аналитики позволяет четко описывать и разграничивать роли заинтересованных сторон и, таким образом, обеспечивает междисциплинарное общение и сотрудничество.

4 Помогает в концептуальном дизайне проектов анализа данных, в первую очередь на ранних этапах

5 Трудно реализовать как масштабируемую платформу вместе с развитием всего проекта

#### 4.16. AI Ops

Джон Томас представляет систематический подход к «операционализации» науки о данных [78], который охватывает управление полным непрерывным жизненным циклом науки о данных. Автор ссылается на этот набор соображений как на AI Ops, который включает: а) Объем б) Понимание с) Сборка (разработка) д) Развертывание и запуск (QA) и е) Развертывание, запуск, управление (Prod)

Для каждого этапа AI-Ops определяет необходимые роли: специалист по данным, бизнес-пользователь, распорядитель данных, поставщик данных, потребитель данных, инженер данных, инженер-программист, операции AI и т. д., Что помогает организовать проект и улучшает координацию. Однако нет никаких указаний относительно того, как эти роли взаимодействуют и взаимодействуют друг с другом.

На этапе охвата методология настаивает на наличии четких бизнес-КПЭ. Работа без четких KPI может привести к тому, что команда будет оценивать успех проекта по производительности модели, а не по ее влиянию на

бизнес. В этом смысле сопоставление показателей производительности модели и элементов доверия / прозрачности с бизнес-ключевыми показателями эффективности действительно является частой проблемой. Без этой информации бизнесу сложно понять, успешен ли проект.

Фаза понимания данных имеет решающее значение: AI-Ops предлагает установить соответствующие правила и политики для управления доступом к данным. Например, это дает возможность командам по анализу данных покупать данные в центральном каталоге: сначала изучать и понимать данные, после чего они могут запросить поток данных и использовать его на этапе сборки.

Эта методология разделяет этап развертывания на два отдельных этапа: оценка качества (QA) и производство (prod), чтобы убедиться, что любая производственная модель соответствует бизнес-требованиям, а также стандартам качества. В целом эта методология ориентирована на реализацию проекта, поэтому большее значение придается развертыванию, инвестициям в ИТ-инфраструктуру и этапам непрерывной разработки и интеграции.

**4** В первую очередь сосредоточены на развертывании и эксплуатации - реализация проекта

**5** Не предлагает руководящих принципов того, как разные роли сочетаются лаборатории и общение друг с другом. Проблемы воспроизводимости и сохранения знаний остаются неисследованными на этапе построения модели

#### 4.17. Рабочий процесс Data Science

Рабочий процесс Data Science [79] Филиппа Гуо представляет современный рабочий процесс исследовательского программирования. Есть четыре основных этапа: подготовка данных, передование между проведением анализа и размышлением для интерпретации результатов, и, наконец, распространение результатов.

На этапе подготовки специалисты по данным собирают данные и очищают их. Приобретение может быть сложной задачей с точки зрения управления данными, происхождения и хранения данных, но определение этого предварительного этапа довольно простое. Отсюда следует основная деятельность науки о данных, то есть фаза анализа. Здесь Гуо представляет отдельный итерационный процесс, в котором готовятся и выполняются сценарии программирования. Результаты этих сценариев проверяются, и после задачи отладки сценарии редактируются. Гуо утверждает, что чем быстрее специалист по данным сможет пройти каждую итерацию, тем больше информации может быть получено за единицу времени.

В то время как фаза анализа включает в себя программирование, фаза размышления включает в себя обдумывание и обсуждение результатов анализа. После проверки набора выходных файлов специалист по данным может делать заметки, проводить встречи и делать сравнения. Понимание, полученное на этом этапе размышлений, используется для изучения новых альтернатив путем корректировки кода сценария и параметров выполнения.

Заключительный этап - распространение результатов, чаще всего в форме письменных отчетов. Задача здесь состоит в том, чтобы взять все различные заметки, наброски, сценарии и файлы выходных данных, созданные в процессе, чтобы помочь в написании.

Некоторые специалисты по данным также распространяют свое программное обеспечение, чтобы другие исследователи могли воспроизводить свои эксперименты.

В целом, отличия анализа от фазы отражения - главная сила этого рабочего процесса Гуо. Исследовательский характер проектов по науке о данных затрудняет создание и разработку водопада. Следовательно, обычно необходимо ходить взад и вперед, чтобы найти подходящие идеи и оптимальную модель для бизнеса. Отделение чистых задач программирования (анализа) от обсуждения результатов очень ценно для специалистов по данным, которые могут потеряться в спирали программирования.

**4** Отличает фазу анализа от фазы отражения

**5** Явно сосредоточен на исследованиях / науке, а не сразу применимо в бизнес-кейсе

#### 4.18. Жизненный цикл EMC Data Analytics

EMC Data Analytics Lifecycle от EMC [80] - это структура для проектов в области науки о данных, которая была разработана для отражения нескольких ключевых моментов: А) Проекты в области науки о данных являются итеративными В) Постоянно проверять, достаточно ли проделана команда для продвижения вперед С) Сосредоточенность работы на обоих впереди и в конце проектов

Эта методология определяет ключевые роли для успешного аналитического проекта: бизнес-пользователь, спонсор проекта, менеджер проекта, аналитик бизнес-аналитики, администратор базы данных, инженер данных, специалист по данным. EMC также включает шесть итерационных этапов, которые показаны в круговой диаграмме: обнаружение данных, подготовка данных, планирование модели, построение модели, передача результатов и ввод в эксплуатацию.

На этапе открытия бизнес-проблема оформляется как задача аналитики, и формулируются исходные гипотезы. Команда также оценивает доступные ресурсы (люди, технологии и данные). Когда имеется достаточно информации для составления аналитического плана, данные преобразуются для дальнейшего анализа на этапе подготовки данных.

EMC отделяет планирование модели от этапа построения модели. Во время планирования модели группа определяет методы, приемы и рабочий процесс, которым она намерена следовать на последующем этапе построения модели. Команда исследует данные, чтобы узнать о взаимосвязях между переменными, а затем выбирает ключевые переменные и наиболее подходящие модели. Отделение планирования модели от этапа построения - это разумный шаг, поскольку он помогает итеративному процессу поиска оптимальной модели. Процесс получения лучшей модели может быть хаотичным и запутанным: на подготовку, программирование и оценку нового эксперимента могут уйти недели. Четкое разделение между планированием модели и построением модели, безусловно, помогает на этом этапе.

Сообщение результатов, представленное как отдельный этап, является отличительной чертой этой методологии. В

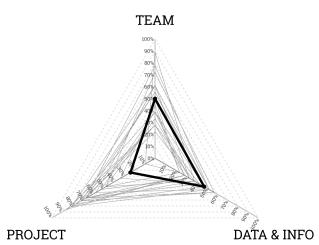


Рисунок 18: Рабочий процесс Data Science

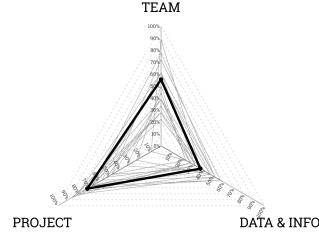


Рисунок 19: Жизненный цикл EMC Data Analytics

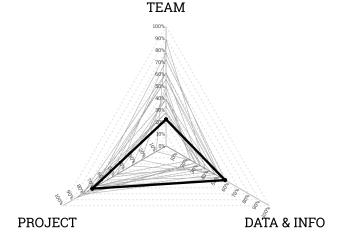


Рисунок 20: На пути к инженерии интеллектуального анализа данных

На этом этапе команда в сотрудничестве с основными заинтересованными сторонами определяет, являются ли результаты проекта успешными или неудачными, основываясь на критериях, разработанных на этапе обнаружения. Команда должна определить ключевые выводы, количественно оценить ценность бизнеса и разработать повествование, чтобы обобщить и передать результаты заинтересованным сторонам. Наконец, на этапе ввода в эксплуатацию команда предоставляет заключительные отчеты, брифинги, код и техническую документацию.

В целом методология EMC дает четкое руководство для жизненного цикла науки о данных и определение ключевых ролей. Он очень обеспокоен тем, что команды чрезмерно сосредотачиваются на этапах со второй по четвертую (подготовка данных, планирование модели и построение модели), и явно не позволяет им приступить к выполнению работы по моделированию до того, как они будут готовы. Несмотря на то, что он устанавливает ключевые роли в проекте по науке о данных, а также то, как они должны сотрудничать и координировать свои действия, в нем не содержится подробностей о том, как команды должны общаться более эффективно.

**4** Предотвращает преждевременные прыжки специалистов по данным в модельную работу

**5** Отсутствует воспроизводимость и управление знаниями настраивать

#### 4.19. К инженерии интеллектуального анализа данных

Marbán et al. [81] предлагают повторно использовать идеи и концепции из процессов моделирования программного обеспечения, чтобы переопределить и дополнить процесс CRISP-DM, а также сделать его стандартом проектирования интеллектуального анализа данных.

Авторы предлагают стандарт, который включает все действия в хорошо организованной манере, описывая процесс поэтапно. Здесь обсуждаются только отличительные моменты. Действия, отсутствующие в CRISP-DM, - это в первую очередь процессы управления проектами, интегральные процессы и организационные процессы. Процессы управления проектом устанавливают структуру проекта и координируют ресурсы проекта на протяжении всего жизненного цикла проекта. CRISP-DM учитывает только план проекта, который является небольшой частью управления проектом. Эта методология включает в себя выбор жизненного цикла и другие процессы, которые не будут объяснены дополнительно: приобретение, поставка, инициирование, планирование проекта, а также мониторинг и контроль проекта.

Интегральные процессы необходимы для успешного завершения проектной деятельности. В эту категорию включены процессы оценки, управления конфигурацией, документации и обучения пользователей.

Организационные процессы помогают достичь более эффективной организации, и рекомендуется адаптировать стандарт SPICE. В эту группу входят процессы улучшения (сбор лучших практик, методов и инструментов), инфраструктура (создание лучших сред) и обучение.

Основные процессы в проектах Data Science Engineering - это процессы разработки, которые делятся на процессы до, KDD и после разработки. Предлагаемая модель процесса проектирования очень полная и почти охватывает все аспекты для успешного выполнения проекта. Он основан на очень хорошо изученных и разработанных инженерных стандартах и, как следствие, очень подходит для крупных корпораций и групп специалистов по анализу данных. Тем не менее, чрезмерные усилия прилагаются к второстепенным, вспомогательным действиям (организационным, интегральным и проектным процессам), а не к основным мероприятиям по развитию. Это дополнительное обязательство может привести к нежелательным эффектам, затрудняющим выполнение проекта и создавая неэффективность из-за сложности процессов. В общем,

**4** Модель тщательного проектирования для успешного выполнения проекта

**5** Проблемы, связанные с командой, остаются неизученными

## 5. Обсуждение

На данный момент проблемы, которые возникают при выполнении проектов по науке о данных, были собраны в разделе 2, и был проведен критический обзор методологий, кратко изложенный в разделе 4. На основе рассмотренных методологий предлагается таксономия методологий, как показано на рисунке 21 год

Критерии такой таксономии сосредоточены на обобщении категорий, адекватно охватываемых каждой методологией.

Например, методология жизненного цикла анализа данных EMC [80] устойчива к проблемам управления проектами и командой, тогда как MIDST [70] явно склоняется в сторону команды

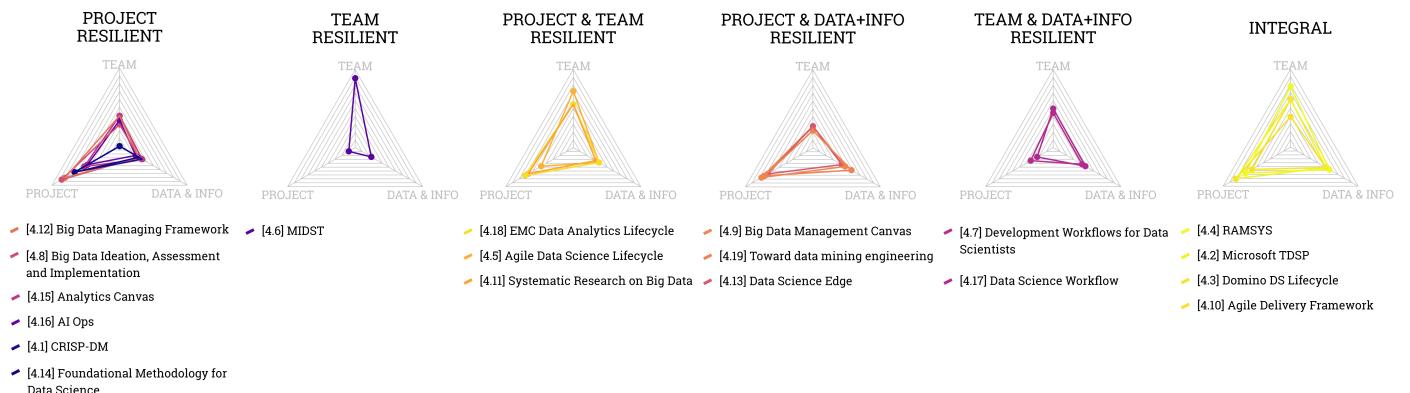


Рисунок 21: Таксономия методологий для проектов в области науки о данных

управление. Эта таксономия предназначена для лучшего понимания типов методологий, доступных для выполнения проектов в области науки о данных.

Помимо отсутствия использования методологии в реальных проектах по науке о данных, что было упомянуто в разделе 2, в литературе отсутствуют полные или целостные методологии. В этом смысле из 19 рассмотренных методологий только 4 относятся к «интегральным». Среди этих «интегральных» методологий некоторые аспекты могут быть улучшены: TSDP [65] сильно зависит от инструментов и технологий Microsoft, и их роли слишком тесно связаны с сервисами Microsoft. Например, роль специалиста по обработке данных TDSP ограничивается клонированием нескольких репозиториев и простым выполнением проекта по науке о данных. Роль специалиста по данным следует разделить на более осозаемые роли и подробно описать их обязанности за пределами вселенной Microsoft.

В жизненном цикле DominoLab [59] отсутствуют тесты для проверки ограничений и качества данных, и даже несмотря на то, что он объединяет групповой подход, он не описывает, как команды должны работать, чтобы эффективно общаться, координировать и сотрудничать. RAMSYS [68] не полностью обеспечивает воспроизводимость и отслеживаемость, а инвестиции в ИТ-ресурсы не полностью учитываются. Наконец, Agile Data Science Lifecycle [69] не слишком обеспокоен ограничениями методов машинного обучения, безопасности данных и конфиденциальности, а также не предлагает методов оценки результатов.

Кроме того, другие методологии, ориентированные на управление проектами, как правило, забывают о команде, выполняющей проект, и часто оставляют позади управление данными и информацией. Что необходимо, так это методология управления, которая учитывает различные потребности науки о данных, ориентированных на данные, а также учитывает ориентированное на приложения использование моделей и других артефактов, созданных в течение жизненного цикла науки о данных, как предложено в [82]. Принимая во внимание рассмотренные методологии, мы думаем, что, поскольку каждый проект в области науки о данных имеет свои особенности и трудно охватить все возможности, было бы рекомендовано заложить основу для построения целостной методологии. Следовательно, мы предлагаем концептуальную

фреймворк, который включает в себя общие черты, которые могла бы иметь интегральная методология управления проектами в области науки о данных. Фреймворк, который мог бы охватить все перечисленные выше подводные камни и проблемы, после того как он был снабжен набором процессов и передового опыта из промышленных, деловых и академических проектов. Эта структура может использоваться другими исследователями в качестве дорожной карты для расширения используемых в настоящее время методологий или для разработки новых. Фактически, любой, у кого есть видение новой методологии, может взять этот шаблон-котел и создать свою собственную методологию.

В этом смысле мы утверждаем, что эффективная методология науки о данных не должна полагаться только на методологии управления проектами и не должна основываться исключительно на методологиях управления командой. Чтобы предложить полное решение для выполнения проектов в области науки о данных, необходимо охватить три области, как это показано на рисунке 22:

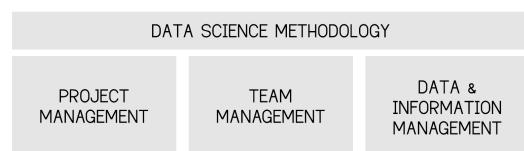


Рисунок 22: Предлагаемые основы интегральных методологий для проектов в области науки о данных

Методологии проекта обычно представляют собой методики, ориентированные на задачи, которые предлагают руководство с шагами, которым нужно следовать, иногда представленные в виде диаграммы. В основном они пытаются определить жизненный цикл науки о данных, который в большинстве случаев очень похож на разные названные методологии [83]: понимание бизнес-проблем, сбор данных, моделирование данных, оценка, реализация. Таким образом, их основная цель - подготовить основные этапы проекта, чтобы его можно было успешно завершить.

Кажется правдой, что проектами в области науки о данных очень трудно управлять из-за неопределенного характера данных среди других факторов, и, следовательно, их процент неудач проектов очень высок. Одним из наиболее важных моментов является то, что в начале любого проекта специалисты по обработке данных не знакомы с данными, и поэтому трудно определить качество данных и их потенциал для достижения определенных бизнес-целей. Это комплиментарно

| УПРАВЛЕНИЕ ПРОЕКТОМ   | УПРАВЛЕНИЕ КОМАНДОЙ  | УПРАВЛЕНИЕ ДАННЫМИ И ИНФОРМАЦИЕЙ                                   |
|---|--|--|
| Определение рабочего процесса жизненного цикла науки о данных | Продвижение и коммуникация научные открытия  | Воспроизведимость: создание хранилища знаний                       |
| Стандартизация структуры папок                                | Определение ролей для облегчения координации между командой                                  | Надежное развертывание: управление версиями кода, данных и моделей |
| Постоянная проектная документация                             | члены и через зaintересованные стороны   | Создание модели для знаний и создания ценности                     |
| Визуализация статуса проекта                                  | Повышение эффективности совместной работы в команде: рабочий процесс git и соглашения о коде | Подкрепление прослеживаемости: превзойти цель                      |
| Согласование науки о данных и бизнес-целей                    |  |  |
| Консолидация показателей эффективности и успеха               |  |  |
| Разделение прототипа и продукта                               |  |  |

Таблица 5: Краткое изложение предложенных принципов интегральных методологий для проектов в области науки о данных

много говорит об определении проекта и постановке целей SMART. Следовательно, наличие эффективной методологии управления проектами является фундаментальным условием успеха любого проекта, особенно для проектов в области науки о данных, поскольку они могут: а) отслеживать, на какой стадии находится проект в любой момент; б) устанавливать цели проекта, его этапы, результаты проекта, его результаты и с) управление объемами задач и сроками поставки.

Что касается управления командой, наука о данных больше не рассматривается как отдельная задача, выполняемая специалистами по данным-единорогам. Это определенно командная работа, которая требует методологии, определяющей способ общения, координации и сотрудничества между членами команды. Вот почему нам нужно добавить еще одно измерение в методологию науки о данных, которое может помочь в управлении ролью и обязанностями каждого члена команды, а также помогает эффективно сообщать о ходе работы команды. Это цель так называемой методологии управления командой. Важным моментом в этом отношении является то, что никакая методология управления проектами сама по себе не может обеспечить успех проекта в области науки о данных. Необходимо разработать методологию совместной работы, которая координирует и управляет задачами, которые должны быть выполнены, и их приоритетом.

Можно сказать, что эти два измерения, управление проектом и управление командой, применимы практически к любому проекту. Конечно, методологии для каждого случая необходимо будет адаптировать и скорректировать. Однако проекты в области науки о данных по своей сути работают с данными, и обычно с большим количеством данных, чтобы извлекать знания, которые могут ответить на некоторые конкретные вопросы бизнеса, машины, системы или агента, тем самым создавая ценность из данных. Поэтому мы считаем принципиально важным иметь общий подход, который помогает генерировать идеи и знания на основе данных, так называемый управление данными и информацией.

Граница между информацией и знаниями может быть тонкой, но, говоря словами [84], информация описывается как уточненные данные, тогда как знания - это полезная информация. Основываясь на пирамиде DIKW, мы можем извлекать информацию из данных, из информационных знаний и из знаний мудрости. Мы утверждаем, что проекты в области науки о данных стремятся не извлекать мудрость из данных, а извлекать из данных полезную информацию, а значит, и знания.

Следовательно, необходимо управлять аналитическими данными и сообщать результаты экспериментов таким образом, чтобы

дань уважения команде и знаниям проекта. В области управления знаниями это означает, что неявные знания должны постоянно преобразовываться в явные. Чтобы осуществить это преобразование от неявного знания, которое является нематериальным, к явному знанию, которое может быть передано в виде товаров, необходимы информация и канал связи. Таким образом, процессы управления информацией и командного взаимодействия полностью взаимосвязаны.

Информация обычно извлекается из данных для получения знаний, и на основе этих знаний принимаются решения. В зависимости от решаемой проблемы, качества и потенциального использования информации для принятия решений и дальнейших действий необходим более или менее человеческий вклад. Основная стратегическая цель аналитического проекта (описательная, диагностическая, прогнозирующая или предписывающая) может повлиять на тип информации, которой необходимо управлять, и, как следствие, также может повлиять на важность этапа управления информацией. Каждая стратегическая цель пытается ответить на разные вопросы, а также определить тип информации, извлекаемой из данных.

Имея только описательную информацию из данных, требуется контролировать и управлять используемыми человеческими знаниями; тогда как с более сложными моделями (диагностическими и прогнозирующими) гораздо важнее управлять данными, моделью, входами, выходами и т. д. Следовательно, с более полезной информацией, извлеченной из данных, путь к решению сокращается.

Кроме того, [85] указывает, что управление знаниями будет ключевым источником конкурентного преимущества для компаний. Поскольку наука о данных растет с увеличением количества алгоритмов и инфраструктуры, способность собирать и использовать уникальные идеи станет ключевым отличием. При сохранении нынешней тенденции через несколько лет наука о данных и применение моделей машинного обучения станут все более популярными. В ходе этой трансформации наука о данных станет меньше заниматься фреймворками, машинным обучением и статистикой, а больше - управлением данными и успешным формулированием бизнес-потребностей, включая управление знаниями.

Учитывая все обстоятельства, таблица 5 суммирует основные основы нашей структуры: принципы изложены так, чтобы можно было проводить дальнейшие исследования для разработки процессов и соответствующих инструментов.

## 6. Выводы

В целом, в этой статье предлагается концептуальная основа для разработки комплексных методологий для управления проектами в области науки о данных. В связи с этим были представлены три основы для таких методологий: проект, команда и управление данными и информацией. Важно отметить, что эта структура должна постоянно развиваться и улучшаться, чтобы адаптироваться к новым вызовам в науке о данных.

Предлагаемая структура основана на критическом обзоре доступных в настоящее время методологий науки о данных, на основе которых была разработана таксономия методологий. Эта таксономия основана на количественной оценке того, как каждая методология преодолевает проблемы, представленные в разделе 2. В этом отношении полученные баллы оценивались соответствующими исследовательскими группами авторов. Несмотря на то, что эта оценка может содержать некоторую предвзятость, мы полагаем, что она дает первоначальную оценку сильных и слабых сторон каждой методологии. В этом смысле мы хотели бы расширить представленный анализ на экспертов и исследователей, а также на практиков в этой области.

## Благодарности

Это исследование не получало какого-либо специального гранта от финансирующих агентств государственного, коммерческого или некоммерческого секторов.

## Приложение А. Треугольная область

Это приложение содержит процедуру для вычисления площади треугольника, взяв за начало первую точку Ферма. Площадь является важным показателем целостности каждой методологии, поскольку расстояния от первой точки Ферма представляют собой их процентную оценку по управлению проектом, командой и информацией.

Учитывая координаты трех вершин  $A$ ,  $B$ ,  $C$  любого треугольника площадь треугольника определяется как:

$$\Delta = \frac{|A_{\text{икс}}(B_y - C_y) + B_{\text{икс}}(C_y - A_y) + C_{\text{икс}}(A_y - B_y)|}{2} \quad |(A.1)$$

Если начало координат является первой точкой Ферма треугольника, координаты вершин  $A$ ,  $B$ ,  $C$  определяются как:

$$\begin{aligned} A_{\text{икс}} \text{знак равно } a \cdot \text{потому что}(\pi/2); B_{\text{икс}} \text{ Азнак равен } a \cdot \text{грех}(\pi/2); C_{\text{икс}} \\ \text{знак равно } b \cdot \text{потому что}7\pi/6; C_{\text{икс}} \text{ равно } b \cdot \text{грех}(7\pi/6); C_{\text{икс}} \\ \text{знак равно } c \cdot \text{потому что}11\pi/6; C_{\text{икс}} \text{ равно } c \cdot \text{грех}(11\pi/6); \end{aligned} \quad |(A.2)$$

Затем, вставив эти выражения для  $A$ ,  $B$ ,  $C$  координаты в A.1 дает окончательное выражение для площади треугольника. Чтобы достичь этого упрощения, необходимо принять во внимание, что  $\text{потому что}(7\pi/6) = -\text{потому что}(11\pi/6)$ .  $\text{Этакже } \text{грех}(7\pi/6) = \text{грех}(11\pi/6) = -1/2$ .

$$\Delta = \left| \frac{\sqrt{3}}{4} (a \cdot b + a \cdot c + b \cdot c) \right| \quad |(A.3)$$

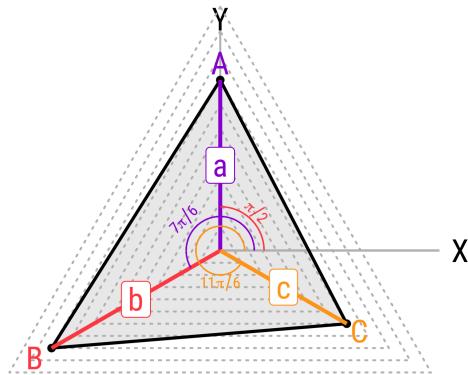


Рисунок А.23: Треугольная диаграмма: начало координат, вершины треугольника и соответствующие углы

## использованная литература

- [1] Натан Бенаич, Ян Хогарт, *State of AI 2019*.
- [2] VentureBeat, Почему 87% проектов в области науки о данных никогда не попадают в производство? (2019).
- [3] New Vantage, NewVantage Partners, опрос руководителей больших данных и искусственного интеллекта, 2019 г. (2019 г.).
- [4] Gartner, Первый шаг к расширенной аналитике.
- [5] Л. Брейман, Статистическое моделирование: две культуры (с комментариями и репликой автора), Статистическая наука 16 (3) (2001) 199–231.
- [6] Л. Гарсия Хименес, Эпистемологический акцент на концепции науки: базовое предложение, основанное на Куне, Поппере, Лакатосе и Фейербенде, *Andamios* 4 (8) (2008) 185–202.
- [7] Дж. Сальц, Потребность в новых процессах, методологиях и инструментах для поддержки групп больших данных и повышения эффективности проектов больших данных, в: Международная конференция IEEE по большим данным (Big Data), 2015 г., стр. 2066–2071. DOI: 10.1109 / Большие данные.2015 г..7363988.
- [8] Дж. Зальц, И. Шамшурин, К. Коннорс, Прогнозирование задач социотехнического выполнения в области науки о данных путем классификации проектов в области науки о данных, Журнал Ассоциации информационных наук и технологий 68 (12) (2017) 2720–2728. DOI: 10.1002 / asi.23873.
- [9] Дж. Зальц, Н. Хотц, Д. Уайлд, К. Стирлинг, Изучение методологий управления проектами, используемых в группах по анализу данных, в: AMCIS, 2018.
- [10] Г. Пятецкий, Crisp-dm, по-прежнему ведущая методология для проектов аналитики, интеллектуального анализа данных или науки о данных, KDD News.
- [11] Джейффи Сальц, Николас Дж. Хотц, CRISP-DM - Управление проектами в области науки о данных (2019).
- [12] В. Гранвиль, Развитие аналитического таланта: становление данными Ученый, 1-е издание, Wiley Publishing, 2014.
- [13] Фрэнк Ло, Что такое наука о данных?
- [14] М. Лукидес, Что такое наука о данных?», « O'Reilly Media, Inc. », 2011 г.
- [15] Д. Конвой, Диаграмма Венна для анализа данных, последнее обращение 13 декабря (2010 г.), 2017 г.
- [16] А. Джонс-Фармер, Р. Хёрл, Чем статистическая инженерия отличается от науки о данных?
- [17] М. Дас, Р. Цуй, Д. Р. Кэмбелл, Г. Агравал, Р. Рамнат, К методам систематического исследования больших данных, в: Международная конференция IEEE по большим данным (Big Data), 2015 г., стр. 2072 –2081. DOI: 10.1109 / BigData.2015 г..7363989.
- [18] Дж. Зальц, Ф. Армор, Р. Шарда, Роли в области науки о данных и типы программ по обработке и анализу данных, Коммуникации Ассоциации информационных систем 43 (1) (2018) 33.
- [19] Джош Уиллс, Определение специалиста по данным (2012).
- [20] П. Уорден, Почему термин «наука о данных» ошибочен, но полезен - O'Reilly Radar (2011).
- [21] Э. Нгаи, Ю. Ху, Ю. Вонг, Ю. Чен, Х. Сунь, Применение методов интеллектуального анализа данных для обнаружения финансового мошенничества: А

- рамки классификации и академический обзор литературы, *Decision Support Systems* 50 (3) (2011) 559–569, о количественных методах обнаружения финансового мошенничества. doi: <https://doi.org/10.1016/j.dss.2010.g..08.006>.
- [22] П.М. Адо, Д. Гуган, Б. Хассани, Анализ кредитного риска с использованием моделей машинного и глубокого обучения, *Риски* 6 (2). DOI: 10.3390 / риски6020038.
- [23] Б. П. Чемберлен, А. Кардосо, Ч. Б. Лю, Р. Паглиари, М. П. Дайзенрот, Прогнозирование ценности времени жизни клиента с использованием встраиваний, *CoRR abs / 1703.02596.arXiv: 1703.02596*.
- [24] Ю. Ким, У. Стрит, Интеллектуальная система для нацеливания на клиентов: подход интеллектуального анализа данных, Системы поддержки принятия решений 37 (2) (2004) 215 - 228. doi: [https://doi.org/10.1016/S0167-9236\(03\)0008-3](https://doi.org/10.1016/S0167-9236(03)0008-3).
- [25] Г. Коу, Ю. Пэн, Г. Ван, Оценка алгоритмов кластеризации для анализа финансовых рисков с использованием методов mcdm, *Информационные науки* 275 (2014) 1–12. doi: <https://doi.org/10.1016/j.ins.2014.g..02.137>.
- [26] Дж. Чен, В. Чен, К. Хуанг, С. Хуанг, А. Чен, Анализ данных финансовых временных рядов с использованием глубоких сверточных нейронных сетей, в: 7-я Международная конференция по облачным вычислениям и большим данным (CCBD), 2016 г. 2016. С. 87–92. DOI: 10.1109 / CCBD.2016 г..027.
- [27] М. де Брюйне, Подходы машинного обучения к анализу медицинских изображений: от обнаружения до диагностики, *Medical Image Analysis* 33 (2016) 94–97, 20-летие журнала *Medical Image Analysis* (Media). doi: <https://doi.org/10.1016/j.media.2016.g..06.032>.
- [28] А. Лавеккья, Подходы машинного обучения к открытию лекарств: методы и приложения, *Drug Discovery Today* 20 (3) (2015). 318 - 331. doi: <https://doi.org/10.1016/j.drudis.2014.g..10.012>.
- [29] П. Ларраньга, Б. Кальво, Р. Сантьяна, К. Бельса, Х. Гальдиано, И. Инза, Х.А. Лозано, Р. Арманьянсас, Дж. Сантафе, А. Перес, В. Роблес, Машинное обучение в биоинформатике, Брифинги по биоинформатике 7 (1) (2006) 86–112. DOI: 10.1093 / bib / bbk007.
- [30] Дж. Ван, С. Тан, Д. Ли, С. Ван, К. Лю, Х. Аббас, А. В. Василакос, Производственное решение для больших данных для активного профилактического обслуживания, *Транзакции IEEE по промышленной информатике* 13 (4) (2017) 2039–2047 гг. DOI: 10.1109 / TII.2017 г..2670505.
- [31] Р. Исмаил, З. Осман, А.А. Бакар, Интеллектуальный анализ данных в производственном планировании и составлении графиков: обзор, в: 2-я конференция по интеллектуальному анализу и оптимизации данных, 2009 г., стр. 154–159. DOI: 10.1109 / DMO.2009 г..5341895.
- [32] Д. Квон, Х. Ким, Дж. Ким, С. Сух, И. Ким, К. Ким, Обзор обнаружения сетевых аномалий на основе глубокого обучения, Кластерные вычисления 22. DOI: 10.1007 / s10586-017-1117-8.
- [33] М. Грбович, В. Радославевич, Н. Джурич, Н. Бхамидипати, Дж. Савла, В. Бхагван, Д. Шарп, Электронная коммерция в вашем почтовом ящике: рекомендации по продукту в масштабе, в: Материалы 21-й Международной конференции ACM SIGKDD по открытию знаний и интеллектуальному анализу данных, KDD '15, Association for Computing Machinery, Нью-Йорк, Нью-Йорк, США, 2015, стр. 1809–1818 гг. DOI: 10.1145/2783258.2788627.
- [34] WX Zhao, S. Li, Y. He, EY Chang, J. Wen, X. Li, Подключение социальных сетей к электронной коммерции: холодный старт рекомендаций по продукту с использованием информации микроблогов, *IEEE Transactions on Knowledge and Data Engineering* 28 (5) (2016) 1147–1159. DOI: 10.1109 / TKDE.2015 г..2508816.
- [35] X. Fang, J. Zhan, Анализ настроений с использованием данных обзора продуктов, *J Big Data* 2. DOI: 10.1186 / s40537-015-0015-2.
- [36] М. Боярски, Д. Д. Теста, Д. Двораковски, Б. Фирнер, Б. Флепп, П. Гоял, Л. Д. Джекель, М. Монфорт, У. Мюллер, Дж. Чжан, X. Zhang, J. Zhao, K. Zieba, End to end learning for беспилотные автомобили, *CoRR abs / 1604.07316.arXiv: 1604.07316*.
- [37] Х. Мин, Искусственный интеллект в управлении цепочкой поставок: теория и приложения, *Международный журнал логистических исследований и приложений* 13 (1) (2010) 13–39. DOI: 10.1080/13675560902736537.
- [38] Дж. Парк, З. Чен, Л. Килиарис, М.Л. Куанг, М.А. Масрур, А.М. Филиппс, Ю.Л. Мерфи, Интеллектуальное управление мощностью транспортного средства на основе машинного обучения оптимальных параметров управления.
- методы и прогнозирование типа дороги и загруженности движения, *IEEE Transactions on Vehicle Technology* 58 (9) (2009) 4741–4756. DOI: 10.1109 / TVT.2009 г..2027710.
- [39] Сью Грин, Достижения в области хранения потоковых данных, анализа в реальном времени и (2019).
- [40] С. Ши, К. Ван, П. Сюй, Х. Чу, Сравнительный анализ современных программных инструментов для глубокого обучения, в: 7-я Международная конференция по облачным вычислениям и большим данным (CCBD), 2016 г., С. 99–104. DOI: 10.1109 / CCBD.2016 г..029.
- [41] Джейкоб Спэльстра, Х. Чжан, Гопи Кумар, Наука о данных не просто происходит, она требует процесса (2016).
- [42] Дж. А. Эспиноза, Ф. Армор, Золотая лихорадка аналитики больших данных: исследовательская основа для координации и управления, в: 49-я Гавайская международная конференция по системным наукам (HICSS), 2016 г., стр. 1112–1121. DOI: 10.1109 / HICSS.2016 г..141.
- [43] М. Колас, И. Финк, Дж. Буват, Р. Намбьяр, Р. Р. Сингх, Решение головоломки с данными: как успешные компании делают большие данные операционными, *Sapgemini Consulting* (2014) 1–18.
- [44] Стеф Карагэуль, Проблемы науки о данных (2018).
- [45] Э. Магуайр, Данные и расширенная аналитика: высокие ставки, высокие награды (2017).
- [46] Дж. Сальц, И. Шамшурина, Методологии командных процессов по работе с большими данными: обзор литературы и определение ключевых факторов успеха проекта, в: Международная конференция IEEE по большим данным (Big Data), 2016 г., стр. 2872–2879. DOI: 10.1109 / BigData.2016 г..7840936.
- [47] А. Кюн, Р. Йоплен, Ф. Рейнхарт, Д. Рёльтген, С. фон Энцберг, Р. Думитреску, Аналитический холст - основа для проектирования и спецификации проектов анализа данных, *Процедура CIRP* 70 (2018) 162–167, 28-я конференция CIRP Design Conference 2018, 23-25 мая 2018 г., Нант, Франция. doi: <https://doi.org/10.1016/j.Procir.2018.g..02.031>.
- [48] Д.К. Беккер, Прогнозирование результатов для проектов больших данных: динамика проектов больших данных (bdpd): исследования в процессе, в: Международная конференция IEEE по большим данным (Big Data), 2017 г., стр. 2320–2330. DOI: 10.1109 / BigData.2017 г..8258186.
- [49] Джоффри Зальц, Николас Дж. Хотц, Недостатки специального управления проектами в области науки о данных (2018).
- [50] Дж. Зальц, И. Шамшурина, К. Коннорс, Структура для описания проектов больших данных, в: В. Абрамович, Р. Альт, Б. Францик (ред.), Семинары по системам бизнес-информации, *Springer International Publishing*, Чам, 2017, с. 183–195.
- [51] Дженифер Прендики, Уроки гибкого машинного обучения от Walmart (2017).
- [52] А.П. Бхардвадж, С. Бхаттахерджи, А. Чаван, А. Дешпанде, А.Дж. Элмор, С. Мэдден, А.Г. Парамесваран, *Datahub*: совместная наука о данных и управление версиями наборов данных в масштабе, *CoRR abs / 1409.0798.arXiv: 1409.0798*.
- [53] К. Бирн, Рабочие процессы разработки для специалистов по данным, *O'Reilly Media*, 2017.
- [54] С. Рэнсботэм, Д. Кирон, П. К. Прентис, Учет пробелов в аналитике, *MIT Sloan Management Review*, 2015.
- [55] Kaggle, Kaggle: ваше сообщество по машинному обучению и науке о данных.
- [56] Ю.Юнг, Мысление о конкуренции: расхождения между Kaggle и реальной наукой о данных (2020).
- [57] М. Ванаэр, К. Бёле, Б. Хеллинграт, Руководство по внедрению больших данных в организациях: методология, основанная на идеях, основанных на бизнесе и данных, и внедрении на основе управления архитектурой предприятия, в: 48-я Гавайская международная конференция, 2015 г. Системные науки, 2015, стр. 908–917. DOI: 10.1109 / HICSS.2015 г..113.
- [58] Э. Колсон, Почему командам по анализу данных нужны универсалы, а не специалисты, *Harvard Business Review*.
- [59] Domino Data Lab, Управление командами Data Science (2017).
- [60] У. Сивараджах, М.М. Камаль, З. Ирани, В. Вираккоди, Критический анализ проблем больших данных и аналитические методы, *Journal of Business Research* 70 (2017) 263–286. doi: <https://doi.org/10.1016/j.jbusres.2016.g..08.001>.
- [61] MN Saunders, C. Rojon, Об атрибуатах критического обзора литературы, *Coaching: An International Journal of The-*

- ogy, Research and Practice 4 (2) (2011) 156–162. DOI: 10.1080/17521882.2011 г..596485.
- [62] С. Браун, Примеры шкалы Лайкера для опросов, Оценка программы ANR, Университет штата Айова, США.
- [63] К. Ширер, Модель crisp-dm: новый план интеллектуального анализа данных, Журнал хранилищ данных 5 (4) (2000) 13–22.
- [64] Н. В. Грейди, Kdd встречает большие данные, в: Международная конференция IEEE по большим данным (Big Data), 2016 г., стр. 1603–1608. DOI: 10.1109 / BigData.2016 г..7840770.
- [65] Microsoft, Team Data Science Process Documentation (2017).
- [66] Скотт Эллис, Структуры, методологии и процессы (2008).
- [67] Николас Дж. Хотц, Джекфри Зальц, Domino Data Science Lifecycle - Управление проектами в области науки о данных (2018).
- [68] С. Мойл, А. Хорхе, Рамис - методология поддержки проектов быстрого удаленного совместного анализа данных, в: ECML / PKDD01 Семинар: Интеграция аспектов интеллектуального анализа данных, поддержки принятия решений и метаобучения (IDDM-2001), том . 64, 2001.
- [69] Р. Джерни, Agile Data Science 2.0: создание полнофункциональных приложений для анализа данных с помощью Spark, O'Reilly Media, Inc., 2017.
- [70] K. Crowston, JS Saltz, A. Rezgui, Y. Hegde, S. You, Социально-технические возможности для стимгергической координации, внедренные в середине, инструмент для групп по анализу данных, Proceedings of the ACM on Human-Computer Interaction 3 ( CSCW) (2019) 1–25.
- [71] Четан Шарма, Ян Овергур, Масштабирование знаний в Airbnb (2016).
- [72] М. Кауфманн, Холст управления большими данными: эталонная модель для создания ценности на основе данных, Большие данные и когнитивные вычисления 3 (1) (2019) 19. DOI: 10.3390 / bdcc3010019.
- [73] Д. Ларсон, В. Чанг, Обзор и будущее направление гибкой разработки, бизнес-аналитики, аналитики и науки о данных, Международный журнал управления информацией 36 (5) (2016) 700–710.
- [74] К. Коллиер, Гибкая аналитика: ориентированный на ценность подход к бизнес-аналитике и хранению данных, Аддисон-Уэсли, 2012.
- [75] Д. Датта, И. Бозе, Управление проектом больших данных: случай с ramco cements limited, Международный журнал экономики производства 165 (2015) 293–306. doi: <https://doi.org/10.1016/j.ijpe.2014.g..12.032>.
- [76] Н. В. Грейди, Дж. А. Пейн, Х. Паркер, Гибкая аналитика больших данных: Analyticsops для науки о данных, в: Международная конференция IEEE по большим данным (Big Data), 2017 г., стр. 2331–2339. DOI: 10.1109 / BigData.2017 г..8258187.
- [77] Джон Роллингс, Основополагающая методология науки о данных (2015).
- [78] Дж. Томас, AI Ops - Управление сквозным жизненным циклом ИИ (2019).
- [79] П.Дж. Гуо, Программные средства для облегчения исследовательского программирования, доктор философии. Диссертация, Стэнфордский университет, Стэнфорд, Калифорния (2012).
- [80] Д. Дитрих, Процессы жизненного цикла аналитики данных, Патент США. 9 262 493 (16 февраля 2016 г.).
- [81] О. Марбан, Дж. Сеговия, Э. Менасалвас, К. Фернандес-Байсан, К инженерии интеллектуального анализа данных: подход к разработке программного обеспечения, Информационные системы 34 (1) (2009) 87–107.
- [82] К. Вальч, Почему гибкие методологии не достигают цели для проектов AI и ML (2020).
- [83] Ф. Форуги, П. Лукш, Методология науки о данных для проектов кибербезопасности, CoRR abs / 1803.04219. arXiv: 1803.04219.
- [84] JC Terra, T. Angeloni, Понимание разницы между управлением информацией и управлением знаниями, KM Advantage (2003) 1–9.
- [85] А. Феррарис, А. Маззолени, А. Девалье, Дж. Кутюрье, Возможности анализа больших данных и управление знаниями: влияние на производительность фирмы, Управленческое решение.