

² Цифровая экономика

³ Искусственный интеллект
⁴ в оценочной деятельности

⁵ Практическое руководство по разработке систем поддержки
⁶ принятия решений оценщиками с использованием языков
⁷ программирования R и Python

⁸ К. А. Мурашев

⁹ 10 сентября 2021 г.

УДК 519(2+8+682)+004.891.2+330.4+338.5

ББК 16.6+22(16+17)+65.25

ГРНТИ 27.43.51+28.23.35+28.23.29+28.23.37+83.03.51

М91

Искусственный интеллект в оценочной деятельности: практическое руководство по разработке систем поддержки принятия решений оценщиками с использованием языков программирования R и Python / К. А. Мурашев — Inkeri, Санкт-Петербург, 12 августа 2021 г. – 10 сентября 2021 г., 88 с.

Данное произведение является результатом интеллектуальной деятельности и объектом авторского права. Распространяется на условиях лицензии [Creative Commons Attribution-Share Alike 4.0 International \(CC BY-SA 4.0\)](#), оригинальный текст которой доступен по [ссылке](#) [5], перевод которого на русский язык доступен по [ссылке](#) [6]. Разрешается копировать, распространять, воспроизводить, исполнять, перерабатывать, исправлять и развивать произведение либо любую его часть в том числе и в коммерческих целях при условии указания авторства и лицензирования производных работ на аналогичных условиях. Все новые произведения, основанные на произведении, распространяемом на условиях данной лицензии, должны распространяться на условиях аналогичной лицензии, следовательно все производные произведения также будет разрешено распространять, изменять, а также использовать любым образом, в т. ч. и в коммерческих целях.

Программный код, разработанный автором и использованный для решения задач, описанных в данном произведении, распространяется на условиях лицензии [Apache License Version 2.0](#) [3], оригинальный текст которой доступен по [ссылке](#) [17], перевод текста которой на русский язык доступен по [ссылке](#) [3]. Программный код на языке R [81], разработанный автором, а также иные рабочие материалы к нему доступны по [ссылке](#) на портале Github [58], а также по [запасной ссылке](#) [59]. Программный код на языке Python [18], разработанный автором, а также иные рабочие материалы к нему доступны по [ссылке](#) на портале Github [60], а также по [запасной ссылке](#) [61].

В процессе разработки данного материала равно как и программного кода автор использовал операционную систему [Kubuntu](#) [12]. Для подготовки данного материала использовался язык \TeX [78] с набором макрорасширений $\text{\LaTeX} 2_{\epsilon}$ [79]. Конкретная техническая реализация заключается в использовании дистрибутива [TexLive](#) [80], редактора \LaTeX [49], компилятора \pdfLaTeX и системы цитирования \BibLaTeX /Biber. Исходный код и дополнительные файлы, необходимые для его компиляции, доступны по [ссылке](#) на портале Github [63], а также по [запасной ссылке](#) [64].

Материал подготовлен в форме гипертекста: ссылки на ресурсы, размещённые в информационно-телекоммуникационной сети «Интернет» [141], выделены синим (blue) цветом, внутренние перекрёстные ссылки выделены красным (red) цветом, библиографические ссылки выделены зелёным (green) цветом. При подготовке данного материала использовался шаблон [KOMAScript Book](#) [43]. В целях облегчения понимания согласования слов в сложноподчинённых предложениях либо их последовательности в тексте реализована графическая разметка, позволяющая понять

структуру предложения: слова, согласованные между собой внутри предложения, подчёркнуты одинаковыми линиями, данное решение применяется только в тех предложениях, в которых, по мнению автора, возможно неоднозначное толкование в части согласования слов внутри него.

Данный материал выпускается в соответствии с философией *Rolling Release* [97], что означает что он будет непрерывно дорабатываться по мере обнаружения ошибок и неточностей, а также в целях улучшения внешнего вида. Идентификатором, предназначенным для определения версии материала, служат её номер и дата релиза, указанные на титульном листе, а также в колонтитулах. История версий приводится в таблице 0.1 на следующей странице-4. Актуальная версия перевода в формате PDF доступна по [ссылке](#) [63], а также по [запасной ссылке](#) [64].

В целях соответствия принципам [устойчивого развития](#) [39, 102], установленным в частности Стратегией [The European Green Deal](#) [67] и являющимся приоритетными для [Единой Европы](#) [33, 15, 89], а также содействия достижению [углеродной нейтральности](#) [83] рекомендуется использовать материал исключительно в электронной форме без распечатывания на бумаге.

Для связи с автором данного перевода можно использовать

- любой клиент, совместимый с протоколом [Tox](#) [70, 103], Tox ID = 2E71 CA29 AF96 DEF6 ABC0 55BA 4314 BCB4 072A 60EC C2B1 0299 04F8 5B26 6673 C31D 8C90 7E19 3B35;
- адрес электронной почты: kirill.murashev@tutanota.de;
- <https://www.facebook.com/murashev.kirill/> [1];

Перед началом работы рекомендуется установить Git (см. подраздел [2.4.1 на с. 74–75](#)), выбрать каталог на локальном компьютере и выполнить следующие команды:

```
$ git clone git@github.com:Kirill-Murashev/  
AI_for_valuers_book.git  
$ git clone git@github.com:Kirill-Murashev/  
AI_for_valuers_R_source.git  
$ git clone git@github.com:Kirill-Murashev/  
AI_for_valuers_Python_source.git
```

История версий

Таблица 0.0.1: История версий материала

№	Номер версии	Дата	Автор	Описание
0	1	2	3	4
1	0.0001.0001	2021-08-14	KAM	Initial

Оглавление

87

88	1. Предисловие	20
89	2. Технологическая основа	28
90	2.1. Параметры использованного оборудования и программного обеспечения	28
91	2.2. Обоснование выбора языков R и Python в качестве средства анализа	
92	данных	28
93	2.2.1. Обоснование отказа от использования табличных процессоров	
94	в качестве средства анализа данных	28
95	2.2.2. R или Python	30
96	2.2.2.1. Общие моменты	30
97	2.2.2.2. Современное состояние	32
98	2.3. Система контроля версий Git	33
99	2.3.1. Общие сведения	33
100	2.3.2. Хеш-функции	36
101	2.3.3. Начало работы с Git и основные команды	39
102	2.3.4. Исключение файлов из списка отслеживания	53
103	2.3.5. Ветки проекта, указатели branch и Head	60
104	2.3.6. Работа с Github	64
105	2.3.6.1. Начало	64
106	2.3.6.2. Настройка соединения с удалённым репозиторием по-	
107	средством протокола SSH	65
108	2.3.6.2.1. Проверка наличия существующих SSH-ключей.	65
109	2.3.6.2.2. Генерация новой пары ключей.	65
110	2.3.6.2.3. Добавление публичного ключа на портал GitHub.	67
111	2.3.6.3. Создание и установка GPG ключа.	68
112	2.3.6.3.1. Основные сведения.	68
113	2.3.6.3.2. Проверка наличия существующих ключей.	69
114	2.3.6.3.3. Генерация пары ключей GPG	69
115	2.3.6.3.4. Добавление публичного ключа на портал GitHub.	70
116	2.3.6.4. Установление связи между локальным и удалённым	
117	репозиториями	71
118	2.3.6.4.1. Отправка содержимого локального репозито-	
119	рия в удалённый.	71

120	2.3.6.4.2. Получение содержимого удалённого репозитория.	72
121	2.3.6.4.3. Обновление репозитория.	72
122	2.3.7. Работа с Git в IDE	73
123	2.3.7.1. Работа в RStudio	73
124	2.3.7.2. Работа в Spyder	73
125	2.3.7.3. Работа в PyCharm	73
126	2.3.8. Заключение	73
127	2.4. Установка и настройка	74
128	2.4.1. Git	74
129	2.4.1.1. Установка на операционных системах, основанных на Debian: Debian, Ubuntu, Mint и т. п.	74
130	2.4.1.2. Установка на операционной системе Windows	74
131	2.4.1.3. Установка на macOS	74
132	2.4.2. R	75
133	2.4.2.1. Установка на операционных системах, основанных на Debian: Debian, Ubuntu, Mint и т. п.	75
134	2.4.2.1.1. Установка на операционных системах Ubuntu, Mint и производных от них.	75
135	2.4.2.1.2. Установка на операционной системе Debian.	76
136	2.4.2.2. Установка на операционных системах Windows и macOS	77
137	2.4.3. RStudio	77
138	2.4.4. Python	78
139	2.4.5. Spyder	78
140	2.4.6. PyCharm	78
141	2.4.7. SQL	78
142	3. Математическая основа анализа данных	79
143	4. Основные понятия	80
144	4.1. Что было раньше: курица или яйцо? Соотношение понятий statistics, machine learning, data mining, artificial intelligence	80
145	5. Начало работы с R	87
146	6. Автоматизированный сбор данных	88

¹⁵² List of Algorithms

Рабочая версия

Список иллюстраций

154	2.3.1.Локальная система контроля версий	34
155	2.3.2.Схема работы централизованной системы контроля версий	35
156	2.3.3.Схема работы распределённой системы контроля версий	36
157	2.3.4.Общая схема работы Git	37
158	2.3.5.Пример вычисления хеша	38
159	2.3.6.Схема состояний файлов в системе Git	42
160	2.3.7.Схема работы указателя Head	60
161	2.3.8.Схема указателей Head и Branch	61
162	2.3.9.Состояние репозитория после переноса указателя Head на ветку Develop	62
163	2.3.10Состояние репозитория при наличии нескольких веток	63

Список таблиц

165	0.0.1 История версий материала	4
166	0.0.2.	19
167	2.1.1.Параметры использованного оборудования	28
168	2.1.2.Параметры использованного программного обеспечения	29
169	4.1.1.Перечень понятий, описывающих группы методов анализа данных .	80
170	4.1.2.Обобщение сведений	86

Список литературы

- [1] URL: <https://www.facebook.com/murashev.kirill/> (дата обр. 28.07.2021).
- [2] Royal Institution Surveyors of Chartered (RICS). *RICS Valuation — Global Standards*. English. UK, London: RICS, 28 нояб. 2019. URL: <https://www.rics.org/eu/upholding-professional-standards/sector-standards/valuation/red-book/red-book-global/> (дата обр. 10.06.2020).
- [3] *Apache 2.0*. URL: http://licenseit.ru/wiki/index.php/Apache_License_version_2.0#.D0.A2.D0.B5.D0.BA.D1.81.D1.82_.D0.BB.D0.B8.D1.86.D0.B5.D0.BD.D0.B7.D0.B8.D0.B8 (дата обр. 17.08.2021).
- [4] Scott Chacon. *Pro Git book*. Перевод на русский язык. URL: <https://git-scm.com/book/ru/v2> (дата обр. 25.08.2021).
- [5] Creative Commons. *Creative Commons Attribution-ShareAlike 4.0 International*. нояб. 2013. URL: <https://creativecommons.org/licenses/by-sa/4.0/legalcode>.
- [6] Creative Commons. *Creative Commons Attribution-ShareAlike 4.0 International RUS*. нояб. 2013. URL: <https://creativecommons.org/licenses/by-sa/4.0/legalcode.ru>.
- [7] *Constitution of the Federal Republic of Germany*. Английский, немецкий. URL: <https://www.bmi.bund.de/EN/topics/constitution/constitutional-issues/constitutional-issues.html> (дата обр. 06.09.2021).
- [8] Microsoft Corporation. *Microsoft Excel*. Английский. URL: <https://www.microsoft.com/en-us/microsoft-365/excel> (дата обр. 20.08.2021).
- [9] CorVVin. *Хеш-функция, что это такое?* URL: <https://habr.com/en/post/534596/> (дата обр. 25.08.2021).
- [10] *Debian official site*. URL: <https://www.debian.org> (дата обр. 08.09.2021).
- [11] *Debian Packages of R Software*. URL: <https://cran.r-project.org/bin/linux/debian/> (дата обр. 08.09.2021).
- [12] Kubuntu devs. *Kubuntu official site*. Kubuntu devs. URL: <https://kubuntu.org/> (дата обр. 17.08.2021).
- [13] KDE e.V. *Plasma. KDE community*. Английский. KDE e.V. URL: <https://kde.org/plasma-desktop/> (дата обр. 19.08.2021).

- [14] Ed25519. URL: <https://ed25519.cr.yp.to/> (дата обр. 04.09.2021).
- [15] Institute Greater for a Europe. *Institute for a Greater Europe official site*. URL: <https://www.institutegreatereurope.com/> (дата обр. 15.04.2021).
- [16] StatSoft Europe. *Statistica: official site*. URL: <https://www.statistica.com/en/> (дата обр. 24.08.2021).
- [17] Apache Software Foundation. *Apache License Version 2.0*. Английский. URL: <https://www.apache.org/licenses/LICENSE-2.0> (дата обр. 17.08.2021).
- [18] Python Software Foundation. Английский. Python Software Foundation. URL: <https://www.python.org/> (дата обр. 17.08.2021).
- [19] The Apache Software Foundation. *OpenOffice Calc*. URL: <https://www.openoffice.org/product/calc.html> (дата обр. 20.08.2021).
- [20] The Document Foundation. *LibreOffice Calc*. Английский. URL: <https://www.libreoffice.org/discover/calc/> (дата обр. 20.08.2021).
- [21] The IFRS Foundation. *IFRS 13 Fair Value Measurement*. UK, London: The IFRS Foundation, 31 янв. 2016. URL: <http://eifrs.ifrs.org/eifrs/bnstandards/en/IFRS13.pdf> (дата обр. 10.06.2020).
- [22] Geeksforgeeks. *Difference between RSA algorithm and DSA*. URL: <https://www.geeksforgeeks.org/difference-between-rsa-algorithm-and-dsa/> (дата обр. 04.09.2021).
- [23] GeoGebra official site. URL: <https://www.geogebra.org/> (дата обр. 26.08.2021).
- [24] Git Download for Windows. URL: <https://git-scm.com/download/win> (дата обр. 29.08.2021).
- [25] Git install on macOS. URL: <https://git-scm.com/download/mac> (дата обр. 29.08.2021).
- [26] Git official site. URL: <https://git-scm.com/> (дата обр. 19.08.2021).
- [27] Git на сервере — Протоколы. URL: <https://git-scm.com/book/ru/v2/Git-%D0%BD%D0%B0-%D1%81%D0%B5%D1%80%D0%B2%D0%B5%D1%80%D0%B5-%D0%9F%D1%80%D0%BE%D1%82%D0%BE%D0%BA%D0%BE%D0%BB%D1%8B> (дата обр. 03.09.2021).
- [28] GitHub Desktop. URL: <https://desktop.github.com/> (дата обр. 19.08.2021).
- [29] Github site. URL: <https://github.com/> (дата обр. 03.09.2021).
- [30] GNU Privacy Guard Howto. URL: <https://help.ubuntu.com/community/GnuPrivacyGuardHowto> (дата обр. 06.09.2021).
- [31] GnuPG — The Universal Crypto Engine. URL: <https://gnupg.org/software/index.html> (дата обр. 06.09.2021).
- [32] Google. *Google Sheets*. URL: <https://www.google.com/sheets/about/> (дата обр. 20.08.2021).

- [33] Lisbon-Vladivostok Work group. *Initiative Lisbon-Vladivostok*. URL: <https://lisbon-vladivostok.pro/> (дата обр. 15.04.2021).
- [34] Homebrew. URL: <https://brew.sh/> (дата обр. 29.08.2021).
- [35] IBM. *SPSS: official page*. URL: <https://www.ibm.com/products/spss-statistics> (дата обр. 24.08.2021).
- [36] IHS Global Inc. *Eviews: official site*. URL: <https://www.eviews.com/home.html> (дата обр. 24.08.2021).
- [37] SAS Institute Inc. *SAS: official site*. URL: https://www.sas.com/en_us/home.html (дата обр. 24.08.2021).
- [38] Intel. *Процессор Intel® Core™ i7-7500U*. Русский. тех. отч. URL: <https://ark.intel.com/content/www/ru/ru/ark/products/95451/intel-core-i7-7500u-processor-4m-cache-up-to-3-50-ghz.html> (дата обр. 19.08.2021).
- [39] Investopedia. *Sustainability*. URL: <https://www.investopedia.com/terms/s/sustainability.asp> (дата обр. 15.04.2021).
- [40] ISO. *Office Open XML*. URL: https://standards.iso.org/ittf/PubliclyAvailableStandards/c071692_ISO_IEC_29500-4_2016.zip (дата обр. 20.08.2021).
- [41] ISO/IEC. *ISO/IEC 10746-2:2009. Information technology "— Open distributed processing "— Reference model: Foundations — Part 2*. English. под ред. ISO/IEC. Standard. ISO/IEC, 15 дек. 2009. URL: <http://docs.cntd.ru/document/431871894> (дата обр. 01.03.2021).
- [42] ISO/IEC. *ISO/IEC 2382:2015. Information technology — Vocabulary*. English. под ред. ISO/IEC. ISO/IEC, 2015. URL: <https://www.iso.org/obp/ui/#iso:std:iso-iec:2382:ed-1:v1:en> (дата обр. 01.03.2021).
- [43] Markus Kohm. *koma-script — A bundle of versatile classes and packages*. 1994–2020. URL: <https://ctan.org/pkg/koma-script> (дата обр. 28.01.2021).
- [44] *LaTeXDraw official page*. URL: <http://latexdraw.sourceforge.net/> (дата обр. 26.08.2021).
- [45] Licenseit.ru. *GNU General Public License*. URL: http://licenseit.ru/wiki/index.php/GNU_General_Public_License (дата обр. 23.08.2021).
- [46] Licenseit.ru. *GNU General Public License version 2*. URL: http://licenseit.ru/wiki/index.php/GNU_General_Public_License_version_2 (дата обр. 23.08.2021).
- [47] Licenseit.ru. *Python License version 2.1*. URL: http://licenseit.ru/wiki/index.php/Python_License_version_2.1 (дата обр. 23.08.2021).
- [48] StataCorp LLC. *Stata: official site*. URL: <https://www.stata.com/> (дата обр. 24.08.2021).
- [49] *LyX official site*. URL: <https://www.lyx.org/> (дата обр. 28.01.2021).

- [50] Machinelearning.ru. *Интеллектуальный анализ данных*. URL: http://www.machinelearning.ru/wiki/index.php?title=%D0%98%D0%BD%D1%82%D0%B5%D0%BB%D0%BB%D0%B5%D0%BA%D1%82%D1%83%D0%B0%D0%BB%D1%8C%D0%BD%D1%8B%D0%B9_%D0%B0%D0%BD%D0%B0%D0%BB%D0%B8%D0%B7_%D0%B4%D0%B0%D0%BD%D0%BD%D1%8B%D1%85 (дата обр. 09.09.2021).
- [51] Machinelearning.ru. *Машинное обучение*. URL: http://www.machinelearning.ru/wiki/index.php?title=%D0%9C%D0%B0%D1%88%D0%B8%D0%BD%D0%BD%D0%BE%D0%B5_%D0%BE%D0%B1%D1%83%D1%87%D0%B5%D0%BD%D0%B8%D0%B5 (дата обр. 09.09.2021).
- [52] Machinelearning.ru. *Нормальное распределение*. URL: http://www.machinelearning.ru/wiki/index.php?title=%D0%9D%D0%BE%D1%80%D0%BC%D0%B0%D0%BB%D1%8C%D0%BD%D0%BE%D0%B5_%D1%80%D0%B0%D1%81%D0%BF%D1%80%D0%B5%D0%B4%D0%B5%D0%BB%D0%B5%D0%BD%D0%B8%D0%B5 (дата обр. 02.03.2021).
- [53] Machinelearning.ru. *Параметрические статистические тесты*. URL: http://www.machinelearning.ru/wiki/index.php?title=%D0%9A%D0%B0%D1%82%D0%B5%D0%B3%D0%BE%D1%80%D0%B8%D1%8F:%D0%9F%D0%B0%D1%80%D0%B0%D0%BC%D0%B5%D1%82%D1%80%D0%B8%D1%87%D0%B5%D1%81%D0%BA%D0%B8%D0%B5_%D1%81%D1%82%D0%B0%D1%82%D0%B8%D1%81%D1%82%D0%B8%D1%87%D0%B5%D1%81%D0%BA%D0%B8%D0%B5_%D1%82%D0%B5%D1%81%D1%82%D1%8B (дата обр. 02.03.2021).
- [54] Machinelearning.ru. *Переобучение*. URL: <http://www.machinelearning.ru/wiki/index.php?title=%D0%9F%D0%B5%D1%80%D0%B5%D0%BE%D0%B1%D1%83%D1%87%D0%B5%D0%BD%D0%B8%D0%B5> (дата обр. 09.09.2021).
- [55] Machinelearning.ru. *Теория сложности вычислений*. URL: http://www.machinelearning.ru/wiki/index.php?title=%D0%92%D1%8B%D1%87%D0%B8%D1%81%D0%BB%D0%B8%D1%82%D0%B5%D0%BB%D1%8C%D0%BD%D0%B0%D1%8F_%D1%81%D0%BB%D0%BE%D0%B6%D0%BD%D0%BE%D1%81%D1%82%D1%8C (дата обр. 09.09.2021).
- [56] Machinelearning.ru. *Искусственный интеллект*. URL: http://www.machinelearning.ru/wiki/index.php?title=%D0%98%D1%81%D0%BA%D1%83%D1%81%D1%81%D1%82%D0%B2%D0%B5%D0%BD%D0%BD%D1%8B%D0%B9_%D0%B8%D0%BD%D1%82%D0%B5%D0%BB%D0%BB%D0%B5%D0%BA%D1%82 (дата обр. 09.09.2021).
- [57] LLC Minitab. *Minitab: official site*. URL: <https://www.minitab.com/en-us/> (дата обр. 24.08.2021).
- [58] Kirill A. Murashev. R. URL: https://github.com/Kirill-Murashev/AI_for_valuers_R_source.
- [59] Kirill A. Murashev. R. URL: <https://web.tresorit.com/l/1Zgvt#kBA5FiY0Qtverp8Rjz6gyg>.
- [60] Kirill A. Murashev. R. URL: https://github.com/Kirill-Murashev/AI_for_valuers_Python_source.
- [61] Kirill A. Murashev. R. URL: <https://web.tresorit.com/l/VGZE5#XqySAkmjYODAIcOp1ZWPmg>.

- [62] Kirill A. Murashev. *RICS Valuation — Global Standards 2020. Russian translation.* TeX. 28 июля 2021. URL: <https://web.tresorit.com/l/oFpJF#xr3UGoxLvsszn4vAaHtjqw>.
- [63] Kirill A. Murashev. *Искусственный интеллект в оценочной деятельности: практическое руководство по разработке систем поддержки принятия решений оценщиками с использованием языков программирования R и Python.* Inkeri. URL: https://github.com/Kirill-Murashev/AI_for_valuers_book.
- [64] Kirill A. Murashev. *Искусственный интеллект в оценочной деятельности: практическое руководство по разработке систем поддержки принятия решений оценщиками с использованием языков программирования R и Python.* Inkeri. URL: https://web.tresorit.com/l/3xiTP#1p8pFnG_9No9izLFd09xaA.
- [65] *Notepad++ site.* URL: <https://notepad-plus-plus.org/> (дата обр. 29.08.2021).
- [66] Linux Kernel Organization. *The Linux Kernel Archives.* Linux Kernel Organization. URL: <https://www.kernel.org/> (дата обр. 26.08.2021).
- [67] European Parliament. *The European Green Deal.* 15 янв. 2020. URL: https://www.europarl.europa.eu/doceo/document/TA-9-2020-0005_EN.html (дата обр. 15.04.2021).
- [68] Philip Zimmermann — creator of PGP. URL: <https://philzimmermann.com/EN/background/index.html> (дата обр. 06.09.2021).
- [69] Risan Bagja Pradana. *Upgrade Your SSH Key to Ed25519.* URL: <https://medium.com/risa/upgrade-your-ssh-key-to-ed25519-c6e8d60d3c54> (дата обр. 04.09.2021).
- [70] Tox Project. *Tox project official site.* URL: <https://tox.chat/> (дата обр. 09.03.2021).
- [71] Qt. Английский. URL: <https://www.qt.io/> (дата обр. 19.08.2021).
- [72] R Foundation. *The Comprehensive R Archive Network.* URL: <https://cran.r-project.org/> (дата обр. 24.08.2021).
- [73] *SHA3-512 online hash function.* URL: https://emn178.github.io/online-tools/sha3_512.html (дата обр. 25.08.2021).
- [74] Stackexchange. *RSA vs. DSA for SSH authentication keys.* URL: <https://security.stackexchange.com/questions/5096/rsa-vs-dsa-for-ssh-authentication-keys> (дата обр. 04.09.2021).
- [75] Stackexchange. *What is the difference between data mining, statistics, machine learning and AI?* URL: <https://stats.stackexchange.com/questions/5026/what-is-the-difference-between-data-mining-statistics-machine-learning-and-ai/21669> (дата обр. 09.09.2021).
- [76] Statsoft. *Solving trees.* URL: <http://statsoft.ru/home/textbook/modules/stclatre.html> (дата обр. 20.08.2021).
- [77] RStudio. *RStudio official site.* Английский. URL: <https://www.rstudio.com/> (дата обр. 19.08.2021).

- [78] CTAN team. *TeX official site*. English. CTAN Team. URL: <https://www.ctan.org/> (дата обр. 15.11.2020).
- [79] LaTeX team. *LaTeX official site*. English. URL: <https://www.latex-project.org/> (дата обр. 15.11.2020).
- [80] *TeXLive official site*. URL: <https://www.tug.org/texlive/> (дата обр. 15.11.2020).
- [81] The R Foundation. *The R Project for Statistical Computing*. Английский. The R Foundation. URL: <https://www.r-project.org/> (дата обр. 17.08.2021).
- [82] Wikipedia. *Bash (Unix shell)*. URL: [https://en.wikipedia.org/wiki/Bash_\(Unix_shell\)](https://en.wikipedia.org/wiki/Bash_(Unix_shell)) (дата обр. 02.09.2021).
- [83] Wikipedia. *Carbon neutrality*. URL: https://en.wikipedia.org/wiki/Carbon_neutrality (дата обр. 15.04.2021).
- [84] Wikipedia. *COVID-19 pandemic*. Английский. URL: https://en.wikipedia.org/wiki/COVID-19_pandemic (дата обр. 18.08.2021).
- [85] Wikipedia. *DSA*. URL: <https://ru.wikipedia.org/wiki/DSA> (дата обр. 04.09.2021).
- [86] Wikipedia. *ECDSA*. URL: <https://ru.wikipedia.org/wiki/ECDSA> (дата обр. 04.09.2021).
- [87] Wikipedia. *Efficient-market hypothesis*. URL: https://en.wikipedia.org/wiki/Efficient-market_hypothesis (дата обр. 29.10.2020).
- [88] Wikipedia. *Euclidean distance*. URL: https://en.wikipedia.org/wiki/Euclidean_distance (дата обр. 18.08.2021).
- [89] Wikipedia. *Greater Europe*. URL: https://en.wikipedia.org/wiki/Greater_Europe (дата обр. 15.04.2021).
- [90] Wikipedia. *HTTPS*. URL: <https://en.wikipedia.org/wiki/HTTPS> (дата обр. 03.09.2021).
- [91] Wikipedia. *Kelly Johnson (engineer)*. URL: [https://en.wikipedia.org/wiki/Kelly%5C_Johnson_\(engineer\)](https://en.wikipedia.org/wiki/Kelly%5C_Johnson_(engineer)) (дата обр. 06.11.2020).
- [92] Wikipedia. *KISS principle*. URL: https://en.wikipedia.org/wiki/KISS_principle (дата обр. 06.11.2020).
- [93] Wikipedia. *List of Linux distributions : Debian — based*. URL: https://en.wikipedia.org/wiki/Category:Debian-based_distributions (дата обр. 26.08.2021).
- [94] Wikipedia. *Office Open XML*. URL: https://ru.wikipedia.org/wiki/Office_Open_XML (дата обр. 20.08.2021).
- [95] Wikipedia. *PGP*. URL: <https://ru.wikipedia.org/wiki/PGP> (дата обр. 06.09.2021).
- [96] Wikipedia. *Robert Gentleman*. URL: [https://en.wikipedia.org/wiki/Robert_Gentleman_\(statistician\)](https://en.wikipedia.org/wiki/Robert_Gentleman_(statistician)) (дата обр. 25.08.2021).

- [97] Wikipedia. *Rolling Release*. URL: https://ru.wikipedia.org/wiki/Rolling_release (дата обр. 28.01.2021).
- [98] Wikipedia. *Rosshaka*. URL: https://en.wikipedia.org/wiki/Ross_Ihaka (дата обр. 25.08.2021).
- [99] Wikipedia. *RSA*. URL: <https://ru.wikipedia.org/wiki/RSA> (дата обр. 04.09.2021).
- [100] Wikipedia. *SHA-3*. URL: <https://ru.wikipedia.org/wiki/SHA-3> (дата обр. 26.08.2021).
- [101] Wikipedia. *SSH*. URL: <https://ru.wikipedia.org/wiki/SSH> (дата обр. 03.09.2021).
- [102] Wikipedia. *Sustainability*. English. URL: <https://en.wikipedia.org/wiki/Sustainability> (дата обр. 15.04.2021).
- [103] Wikipedia. *Wikipedia: Tox protocol*. URL: [https://en.wikipedia.org/wiki/Tox_\(protocol\)](https://en.wikipedia.org/wiki/Tox_(protocol)) (дата обр. 09.03.2021).
- [104] Wikipedia. *Архитектура компьютера*. Russian. URL: https://ru.wikipedia.org/wiki/%D0%90%D1%80%D1%85%D0%B8%D1%82%D0%B5%D0%BA%D1%82%D1%83%D1%80%D0%B0_%D0%BA%D0%BE%D0%BC%D0%BF%D1%8C%D1%8E%D1%82%D0%B5%D1%80%D0%B0 (дата обр. 06.08.2021).
- [105] Wikipedia. *Байесовская вероятность*. URL: https://ru.wikipedia.org/wiki/%D0%91%D0%B0%D0%B9%D0%B5%D1%81%D0%BE%D0%B2%D1%81%D0%BA%D0%B0%D1%8F_%D0%B2%D0%B5%D1%80%D0%BE%D1%8F%D1%82%D0%BD%D0%BE%D1%81%D1%82%D1%8C (дата обр. 09.09.2021).
- [106] Wikipedia. *Высокоуровневый язык программирования*. URL: https://ru.wikipedia.org/wiki/%D0%92%D1%8B%D1%81%D0%BE%D0%BA%D0%BE%D1%83%D1%80%D0%BE%D0%B2%D0%BD%D0%B5%D0%B2%D1%8B%D0%B9_%D1%8F%D0%B7%D1%8B%D0%BA_%D0%BF%D1%80%D0%BE%D0%B3%D1%80%D0%B0%D0%BC%D0%BC%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D1%8F (дата обр. 23.08.2021).
- [107] Wikipedia. *Детерминированный алгоритм*. URL: https://ru.wikipedia.org/wiki/%D0%94%D0%B5%D1%82%D0%B5%D1%80%D0%BC%D0%B8%D0%BD%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%BD%D1%8B%D0%B9_%D0%B0%D0%BB%D0%B3%D0%BE%D1%80%D0%B8%D1%82%D0%BC (дата обр. 25.08.2021).
- [108] Wikipedia. *Дискретное логарифмирование*. URL: https://ru.wikipedia.org/wiki/%D0%94%D0%B8%D1%81%D0%BA%D1%80%D0%B5%D1%82%D0%BD%D0%BE%D0%B5_%D0%BB%D0%BE%D0%B3%D0%B0%D1%80%D0%B8%D1%84%D0%BC%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5 (дата обр. 04.09.2021).
- [109] Wikipedia. *Интегрированная среда разработки*. URL: https://ru.wikipedia.org/wiki/%D0%98%D0%BD%D1%82%D0%B5%D0%B3%D1%80%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%BD%D0%B0%D1%8F_%D1%81%D1%80%D0%B5%D0%B4%D0%B0_%D1%80%D0%B0%D0%B7%D1%80%D0%B0%D0%B1%D0%BE%D1%82%D0%BA%D0%B8 (дата обр. 29.08.2021).

- 432 [110] Wikipedia. *Коллизия хеш-функции*. URL: https://ru.wikipedia.org/wiki/%D0%9A%D0%BE%D0%BB%D0%BB%D0%B8%D0%B7%D0%B8%D1%8F_%D1%85%D0%B5%D1%88-%D1%84%D1%83%D0%BD%D0%BA%D1%86%D0%B8%D0%B8 (дата обр. 25.08.2021).
- 435 [111] Wikipedia. *Конечное поле (поле Галуа)*. URL: https://ru.wikipedia.org/wiki/%D0%9A%D0%BE%D0%BD%D0%B5%D1%87%D0%BD%D0%BE%D0%B5_%D0%BF%D0%BE%D0%BB%D0%B5 (дата обр. 04.09.2021).
- 436 [112] Wikipedia. *Кох, Вернер*. URL: https://ru.wikipedia.org/wiki/%D0%9A%D0%BE%D1%85,_%D0%92%D0%B5%D1%80%D0%BD%D0%B5%D1%80 (дата обр. 06.09.2021).
- 437 [113] Wikipedia. *Непараметрическая статистика*. URL: https://ru.wikipedia.org/wiki/%D0%9D%D0%B5%D0%BF%D0%B0%D1%80%D0%B0%D0%BC%D0%B5%D1%82%D1%80%D0%B8%D1%87%D0%B5%D1%81%D0%BA%D0%B0%D1%8F_%D1%81%D1%82%D0%B0%D1%82%D0%B8%D1%81%D1%82%D0%B8%D0%BA%D0%B0 (дата обр. 20.08.2021).
- 440 [114] Wikipedia. *Переменная (математика)*. URL: https://ru.wikipedia.org/wiki/%D0%9F%D0%B5%D1%80%D0%B5%D0%BC%D0%B5%D0%BD%D0%BD%D0%B0%D1%8F_%D0%B2%D0%B5%D0%BB%D0%B8%D1%87%D0%B8%D0%BD%D0%B0 (дата обр. 20.08.2021).
- 441 [115] Wikipedia. *Переменная (программирование)*. URL: [https://ru.wikipedia.org/wiki/%D0%9F%D0%B5%D1%80%D0%B5%D0%BC%D0%B5%D0%BD%D0%BD%D0%B0%D1%8F_\(%D0%BF%D1%80%D0%BE%D0%B3%D1%80%D0%B0%D0%BC%D0%BC%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5\)](https://ru.wikipedia.org/wiki/%D0%9F%D0%B5%D1%80%D0%B5%D0%BC%D0%B5%D0%BD%D0%BD%D0%B0%D1%8F_(%D0%BF%D1%80%D0%BE%D0%B3%D1%80%D0%B0%D0%BC%D0%BC%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5)) (дата обр. 20.08.2021).
- 442 [116] Wikipedia. *Полнота по Тьюрингу*. URL: https://ru.wikipedia.org/wiki/%D0%9F%D0%BE%D0%BB%D0%BD%D0%BE%D1%82%D0%B0_%D0%BF%D0%BE_%D0%A2%D1%8C%D1%8E%D1%80%D0%B8%D0%BD%D0%B3%D1%83 (дата обр. 23.08.2021).
- 443 [117] Wikipedia. *Принцип Дирихле*. URL: [https://ru.wikipedia.org/wiki/%D0%9F%D1%80%D0%B8%D0%BD%D1%86%D0%B8%D0%BF_%D0%94%D0%B8%D1%80%D0%B8%D1%85%D0%BB%D0%B5_\(%D0%BA%D0%BE%D0%BC%D0%B1%D0%B8%D0%BD%D0%B0%D1%82%D0%BE%D1%80%D0%B8%D0%BA%D0%B0\)](https://ru.wikipedia.org/wiki/%D0%9F%D1%80%D0%B8%D0%BD%D1%86%D0%B8%D0%BF_%D0%94%D0%B8%D1%80%D0%B8%D1%85%D0%BB%D0%B5_(%D0%BA%D0%BE%D0%BC%D0%B1%D0%B8%D0%BD%D0%B0%D1%82%D0%BE%D1%80%D0%B8%D0%BA%D0%B0)) (дата обр. 25.08.2021).
- 444 [118] Wikipedia. *Пятецкий-Шапиро, Илья Иосифович*. URL: https://ru.wikipedia.org/wiki/%D0%9F%D1%8F%D1%82%D0%B5%D1%86%D0%BA%D0%B8%D0%B9-%D0%A8%D0%B0%D0%BF%D0%B8%D1%80%D0%BE,_%D0%98%D0%BB%D1%8C%D1%8F_%D0%98%D0%BE%D1%81%D0%B8%D1%84%D0%BE%D0%B2%D0%B8%D1%87 (дата обр. 09.09.2021).
- 445 [119] Wikipedia. *Расстояние городских кварталов*. URL: https://en.wikipedia.org/wiki/Taxicab_geometry (дата обр. 18.08.2021).
- 446 [120] Wikipedia. *Сверхвысокоуровневый язык программирования*. URL: https://ru.wikipedia.org/wiki/%D0%A1%D0%B2%D0%B5%D1%80%D1%85%D0%B2%D1%8B%D1%81%D0%BE%D0%BA%D0%BE%D1%83%D1%80%D0%BE%D0%B2%D0%BD%D0%B5%D0%B2%D1%8B%D0%B9_%D1%8F%D0%B7%D1%8B%D0%BA_%D0%BF%D1%80%D0%BE%D0%B3%D1%80%D0%B0%D0%BC%D0%BC%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D1%8F (дата обр. 23.08.2021).

- 472 [121] Wikipedia. *Свободная лицензия*. URL: https://ru.wikipedia.org/wiki/%D0%A1%D0%B2%D0%BE%D0%B1%D0%BE%D0%B4%D0%BD%D0%B0%D1%8F_%D0%BB%D0%B8%D1%86%D0%B5%D0%BD%D0%B7%D0%B8%D1%8F (дата обр. 23.08.2021).
- 473
- 474
- 475 [122] Wikipedia. *Свободное программное обеспечение*. Русский. URL: https://ru.wikipedia.org/wiki/%D0%A1%D0%B2%D0%BE%D0%B1%D0%BE%D0%B4%D0%BD%D0%BE%D0%B5_%D0%BF%D1%80%D0%BE%D0%B3%D1%80%D0%B0%D0%BC%D0%BC%D0%BD%D0%BE%D0%B5_%D0%BE%D0%B1%D0%B5%D1%81%D0%BF%D0%B5%D1%87%D0%B5%D0%BD%D0%B8%D0%B5 (дата обр. 18.08.2021).
- 476
- 477
- 478
- 479
- 480 [123] Wikipedia. *Сильная форма Гипотезы эффективного рынка*. URL: https://ru.wikipedia.org/wiki/%D0%93%D0%B8%D0%BF%D0%BE%D1%82%D0%B5%D0%B7%D0%B0_%D1%8D%D1%84%D1%84%D0%B5%D0%BA%D1%82%D0%B8%D0%B2%D0%BD%D0%BE%D0%B3%D0%BE_%D1%80%D1%8B%D0%BD%D0%BA%D0%B0%D0%A2%D1%80%D0%B8_%D1%84%D0%BE%D1%80%D0%BC%D1%8B_%D1%80%D1%8B%D0%BD%D0%BE%D1%87%D0%BD%D0%BE%D0%B9_%D1%8D%D1%84%D1%84%D0%B5%D0%BA%D1%82%D0%B8%D0%B2%D0%BD%D0%BE%D1%81%D1%82%D0%B8 (дата обр. 18.08.2021).
- 481
- 482
- 483
- 484
- 485
- 486
- 487 [124] Wikipedia. *Сценарный язык*. URL: https://ru.wikipedia.org/wiki/%D0%A1%D1%86%D0%B5%D0%BD%D0%B0%D1%80%D0%BD%D1%8B%D0%B9_%D1%8F%D0%B7%D1%8B%D0%BA (дата обр. 23.08.2021).
- 488
- 489
- 490 [125] Wikipedia. *Хеш-функция*. URL: <https://ru.wikipedia.org/wiki/%D0%A5%D0%B5%D1%88-%D1%84%D1%83%D0%BD%D0%BA%D1%86%D0%B8%D1%8F> (дата обр. 25.08.2021).
- 491
- 492
- 493 [126] Wikipedia. *Циммерман, Филипп*. URL: <https://ru.wikipedia.org/wiki/%D0%A6%D0%B8%D0%BC%D0%BC%D0%B5%D1%80%D0%BC%D0%B0%D0%BD,%D0%A4%D0%B8%D0%BB%D0%B8%D0%BF%D0%BF> (дата обр. 06.09.2021).
- 494
- 495
- 496 [127] Wikipedia. *Частотная вероятность*. URL: https://ru.wikipedia.org/wiki/%D0%A7%D0%B0%D1%81%D1%82%D0%BE%D1%82%D0%BD%D0%B0%D1%8F_%D0%B2%D0%B5%D1%80%D0%BE%D1%8F%D1%82%D0%BD%D0%BE%D1%81%D1%82%D1%8C (дата обр. 09.09.2021).
- 497
- 498
- 499
- 500 [128] Wikipedia. *Эллиптическая кривая*. URL: https://ru.wikipedia.org/wiki/%D0%AD%D0%BB%D0%BB%D0%B8%D0%BF%D1%82%D0%B8%D1%87%D0%B5%D1%81%D0%BA%D0%B0%D1%8F_%D0%BA%D1%80%D0%B8%D0%B2%D0%B0%D1%8F (дата обр. 04.09.2021).
- 501
- 502
- 503
- 504 [129] *Xcode page*. URL: <https://developer.apple.com/xcode/> (дата обр. 29.08.2021).
- 505 [130] *Как запустить Bash скрипт в Linux*. URL: <https://wiki.merionet.ru/servernye-resheniya/63/kak-zapustit-bash-skript-v-linux/> (дата обр. 02.09.2021).
- 506
- 507
- 508 [131] *Калькулятор недвижимости*. URL: <https://spb.cian.ru/kalkulator-nedvizhimosti/> (дата обр. 10.09.2021).
- 509

- [132] Кирилл Кринкин. *Введение в архитектуру ЭВМ и элементы ОС. Курс лекций*. Русский. Computer Science Center. URL: <https://www.youtube.com/watch?v=FzN8zzMRTlw&list=PLlb7e2G7aSpRZ9wDzXI-VYpk59acLF0Ir> (дата обр. 23.08.2021).
- [133] Артём Матяшов. *Git. Большой практический выпуск*. Русский. URL: <https://www.youtube.com/watch?v=SEvR780hGtw> (дата обр. 03.09.2021).
- [134] связи и массовых коммуникаций Российской Федерации Министерство цифрового развития. *Свободное программное обеспечение в госорганах*. Русский. URL: <https://www.gnu.org/philosophy/free-sw.ru.html> (дата обр. 18.08.2021).
- [135] Фонд свободного программного обеспечения. *Что такое свободная программа?* Русский. Фонд свободного программного обеспечения. URL: <https://www.gnu.org/philosophy/free-sw.ru.html> (дата обр. 18.08.2021).
- [136] Программирование на C и C++. Онлайн справочник программиста на C и C++. *Оператор*. URL: <http://www.c-cpp.ru/books/operatory> (дата обр. 20.08.2021).
- [137] Виталий Радченко. *Открытый курс машинного обучения. Тема 5. Композиции: бэггинг, случайный лес*. URL: <https://habr.com/en/company/ods/blog/324402/> (дата обр. 20.08.2021).
- [138] Министерство финансов России. *Международный стандарт финансовой отчётности (IFRS) 13 «Оценка справедливой стоимости»*. с изменениями на 11 июля 2016 г. Russian. Russia, Moscow: Минфин России, 28 дек. 2015. URL: <https://normativ.kontur.ru/document?moduleId=1&documentId=326168#10> (дата обр. 10.06.2020).
- [139] Министерство цифрового развития Российской Федерации. *Национальная программа «Цифровая экономика Российской Федерации»*. 29 окт. 2020. URL: <https://digital.gov.ru/ru/activity/directions/858/> (дата обр. 29.10.2020).
- [140] Министерство экономического развития РФ. *Федеральные стандарты оценки*. URL: https://www.consultant.ru/document/cons_doc_LAW_126896/.
- [141] Российская Федерация. *Федеральный Закон «Об информации, информационных технологиях и о защите информации»*. 149-ФЗ. Russian. Russia, Moscow, 14 июля 2006. URL: <https://normativ.kontur.ru/document?moduleId=1&documentId=376603&cwi=22898> (дата обр. 07.07.2020).
- [142] Российская Федерация. *Федеральный закон «Об оценочной деятельности в Российской Федерации»*. 29 июля 1998. URL: <https://normativ.kontur.ru/document?moduleId=1&documentId=396506&cwi=7508> (дата обр. 18.08.2021).

Таблица 0.0.2.

Глава 1.

Предисловие

«Лучший способ в чём-то
разобраться — это попробовать
научить этому другого».
Народная мудрость

«Лучший способ в чём-то
разобраться до конца — это
попробовать научить этому
компьютер».
Дональд Э. Кнут

Целью данной работы является попытка объединения наработок в областях оценочной деятельности и искусственного интеллекта. Автор предпринимает попытку доказать возможность применения современных технологий искусственного интеллекта в сфере оценки имущества, его эффективность и наличие ряда преимуществ относительно иных методов определения стоимости и анализа данных открытых рынков. В условиях заданного руководством России курса на цифровизацию экономики и, в особенности, на развитие технологий искусственного интеллекта [139] внедрение методов машинного обучения в повседневную практику оценщиков представляется логичным и необходимым.

Данная работа писалась в условиях распространения новой коронавирусной инфекции [84], внесшей дополнительный вклад в процессы цифровизации во всём мире. Можно по-разному относиться к проблематике данного явления, однако нельзя отрицать его влияние на общество и технологический уклад ближайшего будущего. Повсеместный переход на технологии искусственного интеллекта, замена человеческого труда машинным, беспрецедентный рост капитализации компаний, сделавших ставку на развитие интеллектуальной собственности, делают невозможным

игнорирование необходимости цифровой трансформации оценочной деятельности в России.

Актуальность предложенного автором исследования заключается во-первых в том, что оно даёт практический инструментарий, позволяющий делать обоснованные, поддающиеся верификации выводы на основе использования исключительно объективных информации и данных,¹ непосредственно наблюдаемых на открытых рынках, без использования каких-либо иных их источников, подверженных субъективному влиянию со стороны их авторов. Во-вторых, предложенные и рассмотренные в данной работе методы обладают весьма широким функционалом, позволяющим использовать их при решении широкого круга задач, выходящих за рамки работы над конкретной оценкой. Важность обеих причин автор видит в том, что на 2021 год в России в сфере оценочной деятельности сложилась ситуация, которую можно охарактеризовать тремя состояниями:

- состояние неопределённости будущего отрасли;
- состояние интеллектуального тупика;
- состояние технологической отсталости.

¹По мнению автора, отличие между информацией и данными заключается в том, что под информацией понимаются:

- знания о предметах, фактах, идеях и т. д., которыми могут обмениваться люди в рамках конкретного контекста [41];
- знания относительно фактов, событий, вещей, идей и понятий, которые в определённом контексте имеют конкретный смысл [42],

таким образом, в контексте данного материала под информацией следует понимать совокупность сведений, образующих логическую схему: теоремы, научные законы, формулы, эмпирические принципы, алгоритмы, методы, законодательные и подзаконные акты и т. п.

Данные же представляют собой:

- формы представления информации, с которыми имеют дело информационные системы и их пользователи [41];
- поддающееся многократной интерпретации представление информации в формализованном виде, пригодном для передачи, связи или обработки [42],

таким образом, в контексте данного материала под данными следует понимать собой совокупность результатов наблюдений о свойствах тех или иных объектов и явлений, выраженных в объективной форме, предполагающей их многократные передачу и обработку.

Например: информацией является знание о том, что для обработки переменных выборки аналогов, имеющих распределение отличное от [нормального](#) [52], в общем случае, некорректно использовать [параметрические методы](#) [53] статистического анализа; данные в этом случае — это непосредственно сама выборка.

Иными словами, оперируя терминологией [архитектуры ЭВМ](#) [104], данные — набор значений переменных, информация — набор инструкций.

Во избежание двусмысленности в тексте данного материала эти термины приводятся именно в тех смыслах, которые описаны выше. В случае необходимости также используется более общий термин «сведения», обобщающий оба вышеуказанных понятия. В ряде случаев, термины используются в соответствии с принятым значением в контексте устоявшихся словосочетаний.

Первая проблема заключается в неопределённости как правового регулирования отрасли, так и её экономики. Введённая около четырёх лет назад система квалификационных аттестатов оценщиков, на которую регулятор, заказчики и, возможно, часть самих оценщиков возлагали надежду как на фильтр, позволяющий оставить в отрасли только квалифицированных специалистов, сократить предложение оценочных услуг и, следовательно, способствовать росту вознаграждений за проведение оценки, не оправдала ожиданий. Несмотря на существенное сокращение числа оценщиков, имеющих право подписывать отчёты об оценке, не произошло никаких значимых изменений ни в части объёма предложения услуг, ни в части уровня цен на них. Фактически произошло лишь дальнейшее развитие уже существовавшего ранее института подписантов отчётов — оценщиков, имеющих необходимые квалификационные документы и выпускающих от своего имени отчёты, в т. ч. и те, в подготовке которых они не принимали участия. В ряде случаев подписант мог и вовсе не читать отчёт либо даже не видеть его в силу своего присутствия в другом регионе, отличном от региона деятельности компании, выпустившей отчёт. При этом, как ни странно, доход таких «специалистов» не вырос существенным образом. Всё это очевидным образом приводит к недовольству регуляторов в адрес оценочного сообщества. В таких условиях следует ожидать неизбежного дальнейшего ужесточения регулирования и усугубления положения добросовестных оценщиков и оценочных компаний. Вместе с тем было бы ошибочным считать, что виной всему являются исключительно сами оценщики и их работодатели. В существенной степени проблемы квалификации и качества работы оценщиков вызваны не их нежеланием добросовестно выполнять свою работу, а отсутствием у заказчиков интереса к серьёзной качественной оценке. Не секрет, что в большинстве случаев оценка является услугой, навязанной требованиями закона либо кредитора, не нужной самому заказчику, которого очевидно волнует не качество отчёта об оценке, а соответствие определённой в нём стоимости ожиданиям и потребностям заказчика, его договорённостям с контрагентами. В таких условиях, с одной стороны, экономика не создаёт спрос на качественную оценку, с другой — сами оценщики не предлагают экономике интересные решения и новые ценности, которые могли бы принести в отрасль дополнительные финансовые потоки.

Вторая проблема тесно связана с первой и выражается в том числе в наблюдаемом на протяжении последних примерно 10 лет падении качества отчётов об оценке и общей примитивизации работы оценщика. Суть данной проблемы можно кратко сформулировать в одной фразе: «раньше молодые оценщики спрашивали „как проанализировать данные рынка и построить модель для оценки“, сейчас они задают вопрос „где взять корректировку на X“». Установление метода корректировок в качестве доминирующего во всех случаях даже без анализа применимости других методов стало логичным итогом процесса деградации качества отчётов об оценке. При этом источником подобных корректировок чаще всего являются отнюдь не данные открытого рынка. Как и в первом случае винить в этом только самих оценщиков было бы неправильным. В условиях работы в зачастую весьма жёстких временных рамках и за небольшое вознаграждение, оценщик часто лишён возможности провести самостоятельный анализ тех или иных свойств открытого рынка, вследствие

и по причине чего вынужден использовать внешние нерыночные данные в том числе и непроверенного качества. Со временем это становится привычкой, убивающей творчество и стремление к поиску истины.

Третья проблема также неразрывно связана с двумя первыми. Отсутствие конкуренции, основанной на стремлении оказывать как можно более качественные услуги, недостаточная капитализация отрасли, выражающаяся в том числе в относительно невысоких зарплатах оценщиков, не вполне последовательное регулирование отрасли со стороны государства — всё это создаёт условия, при которых у оценщиков отсутствует стимул, а зачастую и возможность внедрять инновации.

Данная работа служит следующей основной цели: дать в руки оценщика инструменты, позволяющие ему просто и быстро извлекать полезные сведения из сырых данных открытых рынков, интерпретировать их, выдвигать гипотезы, выбирать среди них наиболее перспективные и в итоге получать готовые модели предсказания различных свойств объекта оценки, в том числе его стоимости. Есть некоторая надежда, что применение технологий искусственного интеллекта позволит, не увеличивая трудоёмкость, а скорее напротив, снижая её, повысить качество работы оценщика, усилить доказательную силу отчётов об оценке и в итоге позволит создать новые ценности, предлагаемые оценщиками экономике, государству, потребителям, а главное всему обществу.

Особенностью данной работы является её практическая направленность: в тексте содержатся все необходимые инструкции, формулы, описания и фрагменты программного кода либо ссылки на них, необходимые и достаточные для воспроизведения всех рассмотренных методов и их описания в отчётах об оценке.

Данная работа состоит из двух частей. Первая посвящена в большей степени теории, описанию методов, а также применению языка [R](#) [81]. Вторая имеет большую практическую направленность и содержит руководства по применению языка [Python](#) [18]. Объяснение данного факта содержится далее в разделе ССЫЛКА. В работе будут рассмотрены следующие вопросы:

- a) автоматизированный сбор данных с веб-ресурсов;
- b) семантический анализ текстов объявлений;
- c) работа с геоданными;
- d) первичная интерпретация и визуализация данных открытых рынков;
- e) проверка статистических гипотез;
- f) задачи классификации;
- g) корреляционный анализ;
- h) регрессионный анализ;
- i) анализ временных рядов;

- 661 j) задачи многомерного шкалирования;
- 662 k) байесовская статистика;
- 663 l) деревья классификации;
- 664 m) случайные леса;
- 665 n) нейронные сети;
- 666 o) глубокое обучение;
- 667 p) обучение с подкреплением;
- 668 q) нечёткая логика.

669 Вышеприведённый перечень не является исчерпывающим и будет дорабатываться
670 по мере развития проекта.

671 Данная работа основана на четырёх основополагающих принципах и предпосыл-
672 ках.

- 673 a) *Принцип «вся информация об активе учтена в его цене».* Данный принцип
674 говорит о том, что существует функциональная зависимость между ценой ак-
675 тива (обязательства) и его свойствами. Он тесно связан с [Гипотезой эффектив-](#)
676 [ного рынка \[87\]](#), лежащей в основе технического биржевого анализа. При этом
677 для целей настоящей работы данная гипотеза принимается в её [сильной форме](#)
678 [эффективности \[123\]](#). С точки зрения оценщика это означает, что нет необхо-
679 димости искать какие-либо данные кроме тех, которые непосредственно и объ-
680 ективно наблюдаются на рынке.
- 681 b) *Принцип «максимального использования релевантных наблюдаемых исход-*
682 *ных данных и минимального использования ненаблюдаемых исходных дан-*
683 *ных».* Данный принцип согласуется с требованиями п. 3 [Международного](#)
684 [стандарта финансовой отчётности 13 «Оценка справедливой стоимости» \[138\]](#)
685 [\(IFRS 13 \[21\]\)](#), а также, например, принципами [Всемирных стандартов оцен-](#)
686 [ки RICS \[62\] \(RICS Valuation — Global Standards \[2\]\)](#) и основывается на них.
687 С точки зрения оценщика данный принцип означает, что лучшая практика
688 оценки заключается в работе непосредственно с данными открытых рынков,
689 а не чьей-либо их интерпретацией, существующей, например, в виде готовых
690 наборов корректировок, порой весьма далёких от реальности.
- 691 c) *Принцип KISS [92] (keep it simple stupid, вариации: keep it short and simple,*
692 *keep it simple and straightforward и т. п.),* предложенный американским авиа-
693 инженером [Келли Джонсоном \[91\]](#), ставший официальным принципом проек-
694 тирования и конструирования ВМС США с 1960 г. Данный принцип заклю-
695 чается в том, что при разработке той или иной системы следует использовать
696 самое простое решение из возможных. Применительно к тематике данной ра-
697 боты это означает, что в тех случаях, когда автор сталкивался с проблемой

698 выбора способа решения задачи в условиях неопределённости преимуществ
699 и недостатков возможных вариантов, он всегда выбирал самый простой спо-
700 соб. Например в задаче кластеризации, выбирая между видами расстояний,
701 автор делает выбор в пользу [евклидова](#) либо [манхэттенского](#) расстояний [88,
702 119].

- 703 d) *Принцип «не дай алгоритму уничтожить здравый смысл».* Данный прин-
704 цип означает необходимость самостоятельного осмысления всех результатов
705 выполнения процедур, в т. ч. и промежуточных. Возможны ситуации, когда
706 полученные результаты могут противоречить здравому смыслу и априорным
707 знаниям о предметной области, которыми обладает оценщик либо пользова-
708 тели его работы. Следует избегать безоговорочного доверия к результатам,
709 выдаваемым алгоритмами. Если построенная модель противоречит априор-
710 ным знаниям об окружающей реальности, то следует помнить, что другой
711 реальности у нас нет, тогда как модель может быть скорректирована либо
712 заменена на другую.

713 Все описанные этапы действий описаны таким образом, что позволяют сразу же
714 без каких-либо дополнительных исследований воспроизвести всё, что было реали-
715 зовано в данной работе. От пользователей потребуется только установить необхо-
716 димые программные средства, создать свой набор данных для анализа и загрузить
717 его в пакет. Все действия по установке и настройке описаны внутри данного руко-
718 водства. Важным аспектом является то обстоятельство, что при подготовке данно-
719 го исследования использовалось исключительно [свободное программное обеспече-](#)
720 [ние](#) [135, 122, 134]. Таким образом, любой читатель сможет воспроизвести все опи-
721 санные действия без каких-либо затрат на приобретение тех или иных программных
722 продуктов.

723 От пользователей данного руководства не требуется наличие специальных позна-
724 ний в области разработки программного обеспечения, software engineering и иных
725 аспектов computer science. Некоторые понятия вроде «класс», «метод», «функция»,
726 «оператор», «регулярные выражения» и т. п. термины из сферы программирова-
727 ния могут встречаться в тексте руководства, однако их понимание либо непонима-
728 ние пользователем не оказывает существенного влияния на восприятие материала
729 в целом. В отдельных случаях, когда понимание термина является существенным,
730 как например в случае с термином «переменная», в тексте руководства приводится
731 подробное объяснение смысла такого термина, доступное для понимания неспеци-
732 алиста.

733 Также от пользователей руководства не требуется (хотя и является желательным)
734 глубокое понимание математической статистики, дифференциальных вычислений,
735 линейной алгебры, комбинаторики, методов исследования операций, методов опти-
736 мизации и иных разделов математики и математической статистики, хотя и пред-
737 полагается наличие таких познаний на уровне материала, включённого в школь-
738 ную программу и программу технических и экономических специальностей вузов
739 России. В тексте руководства приводится описание смысла и техники всех приме-
740 нённых статистических методов, математических операций и вычислений в объёме,

достаточном, по мнению автора, для обеспечения доказательности при использовании методов, рассмотренных в данной работе. Автор всегда приводит ссылки на материалы, подтверждающие приведённые им описания за исключением случаев общеизвестных либо очевидных сведений. Особое внимание автор уделяет соблюдению требований к информации и данным, имеющим существенное значение для определения стоимости объекта оценки, установленных Федеральным законом «Об оценочной деятельности в Российской Федерации» [142], а также Федеральными стандартами оценки [140].

Сведения, приведённые в настоящем руководстве, являются, по мнению автора, достаточными для обеспечения выполнения вышеуказанных требований к информации, содержащейся в отчёте об оценке. Таким образом, использование описаний процедур, приведённых в настоящем руководстве, скорее всего должно быть достаточным при использовании изложенных в нём методик в целях осуществления оценочной деятельности и составлении отчёта об оценке. Однако, автор рекомендует уточнять требования, предъявляемые к отчёту об оценке со стороны саморегулируемой организации, в которой состоит оценщик, а также со стороны заказчиков и регуляторов.

В силу свободного характера лицензии, на условиях которой распространяется данная работа, она, равно как и любая её часть, может быть скопирована, воспроизведена, переработана либо использована любым другим способом любым лицом в т. ч. и в коммерческих целях при условии распространения производных материалов на условиях такой же лицензии. Таким образом, автор рекомендует использовать тексты, приведённые в настоящем руководстве для описания выполненных оценщиком процедур.

По мнению автора, данное руководство и описанные в нём методы могут быть особенно полезны в следующих предметных областях:

- оценка и переоценка залогов и их портфелей;
- контроль за портфелями залогов со стороны регулятора банковской сферы;
- оценка объектов, подлежащих страхованию, и их портфелей со стороны страховщиков;
- оценка объектов со стороны лизинговых компаний;
- оценка больших групп активов внутри холдинговых компаний и предприятий крупного бизнеса;
- мониторинг стоимости государственного и муниципального имущества;
- оценка в целях автоматизированного налогового контроля;
- государственная кадастровая оценка;
- экспертиза отчётов об оценке, контроль за деятельностью оценщиков со стороны СРО.

Иными словами, особенная ценность применения методов искусственного интеллекта в оценке возникает там, где имеет место необходимость максимальной беспристрастности и незаинтересованности в конкретном значении стоимости.

В данном руководстве не содержатся общие выводы касательно параметров открытых рынков как таковых, не выводятся общие формулы, применимые всегда и для всех объектов оценки. Вместо этого в распоряжение пользователей предоставляется набор мощных инструментов, достаточный для моделирования ценообразования на любом открытом рынке, определения стоимости любого объекта оценки на основе его актуальных данных. В случае необходимости пользователь, применяя рассмотренные методы, может самостоятельно разработать предсказательную модель для любых рынков и объектов. Забегая вперёд, можно сказать, что при решении конкретной практической задачи применение всех описанных методов не является обязательным, а если быть точным — явно избыточным. В тексте руководства содержатся рекомендации по выбору методов на основе имеющихся свойств данных, рассматриваются сильные и слабые стороны каждого из них.

Несмотря на изначально кажущуюся сложность и громоздкость методов, при более детальном знакомстве и погружении в проблематику становится ясно, что применение предложенных реализаций методов существенно сокращает время, необходимое для выполнения расчёта относительно других методов сопоставимого качества, а сама процедура сводится к написанию и сохранению нескольких строк кода при первом применении и их вторичному многократному использованию для новых наборов данных при будущих исследованиях.

Автор выражает надежду, что данное руководство станет для кого-то первым шагом на пути изучения языков [R](#) [81] и [Python](#) [18], а также погружения в мир анализа данных, искусственного интеллекта и машинного обучения.

Глава 2.

Технологическая основа

2.1. Параметры использованного оборудования и программного обеспечения

При выполнении всех описанных в данной работе процедур, равно как и написании её текста использовалась следующая конфигурация оборудования.

Таблица 2.1.1. Параметры использованного оборудования

№	Категория	Модель (характеристика)	Источник
0	1	2	3
1	Процессор	4 × {} Intel ® Core ™ i7-7500U CPU @ 2.70GHz	[38]
2	Память	11741076B	

При выполнении всех описанных в данной работе процедур, равно как и написании её текста использовалась следующая конфигурация программного обеспечения.

Как видно из таблиц 2.1, 2.1 для анализа данных и разработки систем поддержки принятия решений на основе искусственного интеллекта вполне достаточно оборудования, обладающего средними характеристиками, а также свободных или, по крайней мере, бесплатных программных средств.

2.2. Обоснование выбора языков R и Python в качестве средства анализа данных

2.2.1. Обоснование отказа от использования табличных процессоров в качестве средства анализа данных

На сегодняшний день очевиден факт того, что доминирующим программным продуктом, используемым в качестве средства выполнения расчётов, в среде русских оценщиков является приложение MS Excel [8]. Следом за ним идут его бесплатные аналоги LibreOffice Calc и OpenOffice Calc [20, 19], первый из которых

Таблица 2.1.2. Параметры использованного программного обеспечения

№	Категория/наименование	Значение/версия	Источник
0	1	2	3
1	Операционная система	Kubuntu 20.04	[12]
2	KDE Plasma	5.18.5	[13]
3	KDE Frameworks	5.68.0	[13]
4	Qt	5.12.8	[71]
5	R	4.1.1 (2021-08-10) "— "Kick Things"	[81]
6	RStudio	1.4.1717	[77]
7	Git	2.25.1	[26]
8	Github Desktop	2.6.3-linux1	[28]
9	Geogebra Classic	6.0.660.0-offline	[23]
10	LaTeXDraw	4.0.3-1	[44]
11	Python	3.8.10	
12	Spyder	3.3.6	
13	PyCharm Community	2021.2.1	
14	Kate	19.12.3	

является также не только бесплатным, но и [свободным программным обеспечением](#) [135, 122, 134]. В ряде случаев используется [Google Sheets](#) [32]. Не оспаривая достоинства этих продуктов, нельзя не сказать о том, что они являются универсальными средствами обработки данных общего назначения и, как любые универсальные средства, сильны своей многофункциональностью и удобством, но не шириной и глубиной проработки всех функций. Во всех вышеуказанных программных продуктах в виде готовых функций реализованы некоторые основные математические и статистические процедуры. Также само собой присутствует возможность выполнения расчётов в виде формул, собираемых вручную из простейших [операторов](#) [136]. Однако возможности этих продуктов для профессионального анализа данных абсолютно недостаточны. Во-первых, в них имеются ограничения на размер и размерность исследуемых данных. Во-вторых, в них отсутствуют средства реализации многих современных методов анализа данных. Если первое ограничение не столь важно для оценщиков, редко имеющих дела с по-настоящему большими наборами данных и существенным числом [переменных](#) [114, 115] в них, второе всё же накладывает непреодолимые ограничения на пределы применимости таких программных продуктов. Например, ни одно из вышеперечисленных приложений не позволяет использовать методы [непараметрической статистики](#) [113] либо, например, решить задачи построения [деревьев классификации](#) [76] и их [случайных лесов](#) [137]. Таким образом, следует признать, что, оставаясь высококачественными универсальными средствами для базовых расчётов, вышеперечисленные приложения не могут быть использованы для профессионального анализа данных на современном уровне.

При этом их использование порой бывает необходимым на первоначальном исследовании

дования. Некоторые исходные данные, предоставляемые оценщику для обработки, содержатся в электронных таблицах. Такие таблицы помимо полезных сведений могут содержать посторонние данные, тексты, графики и изображения. В практике автора был случай предоставления ему для анализа данных в форме электронной таблицы формата [xlsx](#) [94, 40], имеющей размер около 143 МБ, содержащей помимо подлежащей анализу числовой информации о товарах их рекламные описания в текстовом виде и фотографии, составляющие свыше 90 % размера файла. Тем не менее просмотр исходных данных средствами табличных процессоров и создание нового файла, содержащего только необходимые для анализа данные, нередко является подготовительным этапом процесса анализа. В последующих разделах будут даны практические рекомендации касательно его реализации. По мнению автора, по состоянию на 2021 год лучшим табличным процессором является [LibreOffice Calc](#) [20], превосходящий [MS Excel](#) [8] по ряду характеристик.

2.2.2. R или Python

2.2.2.1. Общие моменты

Можно с уверенностью сказать, что по состоянию на второе полугодие 2021 года доминирующими и самыми массовыми техническими средствами анализа данных, машинного обучения и разработки искусственного интеллекта¹ являются языки программирования [R](#) [81] и [Python](#) [18]. Оба они являются [сверхвысокоуровневыми](#) [120] [сценарными](#) (скриптовыми) [124] языками программирования. Высокоуровневым называется такой язык программирования, в основу которого заложена сильная абстракция, т. е. свойство описывать данные и операции над ними таким образом, при котором разработчику не требуется глубокое понимание того, как именно машина их обрабатывает и исполняет [106]. [Сверхвысокоуровневым](#) [120] языком является такой язык программирования, в котором реализована очень сильная абстракция. Иными словами, в отличие от [языков программирования высокого уровня](#) [106], в коде, разработанном на которых, описывается принцип «как нужно сделать», код, выполненный на [сверхвысокоуровневых языках](#) [120] описывает лишь принцип «что нужно сделать». [Сценарным](#) (скриптовым) [124] языком называется такой язык программирования, работа которого основана на исполнении сценариев, т. е. программ, использующих уже готовые компоненты. Таким образом, можно сделать вывод, что [сверхвысокоуровневые языки](#) лучше всего подходят для тех, кто только начинает погружаться в программирование и не обладает экспертными знаниями в вопросах [архитектуры ЭВМ](#) [104].²

Оба языка распространяются на условиях [свободных лицензий](#) [121] с незначительными отличиями. [R](#) распространяется на условиях лицензии [GNU GPL 2](#) [46], [Python](#) — на условиях лицензии [Python Software Foundation License](#) [47], являющейся совместимой с [GNU GPL](#) [45]. Отличия между ними не имеют никакого

¹Разница между этими понятиями будет описана далее в ССЫЛКА

²Для первичного ознакомления с вопросами архитектуры ЭВМ автор рекомендует просмотреть [данный курс лекций](#) [132].

практического значения для целей настоящего руководства и применения любого из этих языков в оценочной деятельности в целом. Следует лишь знать основной факт: использование этих языков является легальным и бесплатным в том числе и для коммерческих целей. Основное отличие между этими языками заключается в частности в том, что Python — язык общего назначения, широко применяемый в различных областях, тогда как R — специализированный язык статистического анализа и машинного обучения. В целом можно сказать, что задачи анализа данных могут одинаково успешно решаться средствами обоих языков. Также они оба являются [Тьюринг-полными](#) [116] языками.

Преимущества R основаны на том факте, что он изначально был разработан двумя профессиональными статистиками: [Ross Ihaka](#) [98], [Robert Gentleman](#) [96], по первым буквам имён которых он и был назван. Дальнейшее развитие языка также осуществляется прежде всего силами профессиональных математиков и статистиков, вследствие чего для R реализовано значительное количество библиотек, выполняющих практически все доступные на сегодняшнем уровне развития науки статистические процедуры. Кроме того, можно быть уверенным в абсолютной корректности всех алгоритмов, реализованных в этих библиотеках. К тому же этот язык особенно популярен в академической среде, что означает факт того, что в случае, например, выхода какой-то статьи, описывающей новый статистический метод, можно быть уверенным, что соответствующая библиотека, реализующая этот метод выйдет в ближайшее время либо уже вышла. Кроме того, важным преимуществом R являются очень хорошо проработанные средства вывода графической интерпретации результатов анализа.

Недостатки R, как это часто бывает, следуют из его достоинств. Язык и его библиотеки поддерживаются в первую очередь силами математиков-статистиков, а не программистов, что приводит к тому, что язык относительно плохо оптимизирован с точки зрения software engineering, многие решения выглядят неочевидными и неоптимальными с точки зрения способов обращения к памяти, интерпретации в машинные команды, исполнения на процессоре. Это приводит к высокому потреблению ресурсов машины, в первую очередь памяти, медленному исполнению процедур. При этом, говоря о медленном исполнении, следует понимать относительность этой медлительности. Выполнение команды за 35 мс вместо 7 мс не замечается человеком и обычно не имеет сколько-нибудь определяющего значения. Проблемы с производительностью становятся заметны только при работе с данными большой размерности: миллионы наблюдений, тысячи переменных. В практических задачах, с которыми сталкиваются оценщики, подобная размерность данных выглядит неправдоподобной, вследствие чего можно говорить об отсутствии существенных недостатков языка R для целей применения в оценочной деятельности в целом и в целях задач, решаемых в данном руководстве, в частности. Следующей условной проблемой R является огромное количество библиотек³ и ещё более огромное количество возможных вариантов решения задач и предлагаемых для этого методов. Даже

³По состоянию на 24 августа 2021 существует 18089 официальных библиотек, содержащихся на [официальной странице](#) [72] проекта.

опытный аналитик может растеряться, узнав о том, что его задача может быть решена десятками способов, выбор лучшего из которых сам по себе является нетривиальной задачей. Данную особенность конечно же нельзя считать недостатком самого языка R.

Преимуществом Python является его универсальность и существенно большая распространённость. Освоение основ данного языка для целей одной предметной области может быть полезным в дальнейшем, если по каким-то причинам оценщик захочет решать с его помощью задачи иного класса. Данный язык разработан и поддерживается профессиональными программистами, что означает его относительно приемлемую оптимизацию, превосходящую R, но уступающую, например C++.

К недостаткам Python можно отнести меньшее число библиотек, содержащих статистические процедуры. Кроме того, нет такой же уверенности в безупречности их алгоритмов. При этом следует отметить, что подобные риски присутствуют лишь в новых библиотеках, реализующих экспериментальные либо экзотические статистические процедуры. Для целей оценки как правило вполне достаточно уже относительно отработанных и проверенных библиотек.

Подводя итог, можно сказать, что нет однозначного ответа, какой из вышеупомянутых языков является предпочтительным для целей анализа данных в оценке. R развивается, оптимизируется и всё больше избавляется от «детских болезней» неоптимизированности, для Python создаются новые мощные библиотеки статистического анализа. Поэтому вопрос остаётся открытым.

Следует кратко упомянуть о том, что помимо R и Python в целях анализа данных также используются вендорские программные продукты такие как SAS [37], SPSS [35], Statistica [16], Minitab [57], Stata [48], EvIEWS [36] и ряд других. Однако все они являются платными, при этом стоимость лицензии на самый мощный из них — SAS начинается, как правило, от нескольких десятков тысяч долларов. В остальном, кроме привычного для большинства пользователей графического интерфейса они не имеют явных преимуществ перед R и Python, предоставляя при этом даже меньше возможностей.

2.2.2.2. Современное состояние

Вышеприведённый текст, содержащийся в предыдущей секции (2.2.2.1) был написан автором в 2019 году. За прошедший период произошли некоторые изменения, требующие внимания. В настоящее время Python серьёзно опережает R по распространённости в среде аналитиков данных. Можно говорить о некотором консенсусе, согласно которому R является средством разработки и анализа данных для научных целей, тогда как Python применяется в бизнес среде. Несмотря на это, автор считает, что в целях анализа данных данные языки вполне взаимозаменяемы. Некоторые библиотеки портированы из одного из них в другой. При этом нельзя не признать, что за последние годы R существенно сдал позиции в пользу Python. В особенности это справедливо именно для российского рынка разработки систем анализа данных. Определённый пик интереса к R в России имел место в 2015–2017 годах, после чего его популярность пошла на спад. В мире пик интереса к R пришёлся на 2016–2018

годы после чего его популярность стабилизировалась. Язык продолжает активно развиваться.

В российской практике коммерческого анализа данных его заказчики, как правило, требуют реализации на Python, применение вместо него R чаще всего приходится обосновывать отдельно. Таким образом, можно говорить о том, что применение Python де факто является стандартом. Кроме того, продвижению Python во всём мире способствует позиция компаний интернет-гигантов, использующих его в своих системах машинного обучения. Следующим фактором успеха Python является его широкое распространение в теме разработки нейронных сетей, также являющееся следствием практик крупных IT-компаний. Также Python широко распространён и за пределами области анализа данных, что означает существенно большее число специалистов, владеющих им. При этом для R разработан ряд уникальных отраслевых библиотек, содержащих специфические функции. R безоговорочно лидирует в области биоинформатики, моделирования химических процессов, социологии.

При этом, R по-прежнему предоставляет существенно более широкие возможности визуализации, а также позволяет легко разрабатывать веб-интерфейсы посредством [Shiny](#). R имеет отличный инструмент написания документации к коду в процессе разработки самого кода — [R Markdown](#).

Подводя итоги, можно сказать о том, что современным оценщикам следует иметь навыки разработки и анализа данных с использованием обоих этих языков: R поможет применять самые свежие методы и создавать качественные понятные пользователям описания и визуализации, Python пригодится там, где требуется разработка серьёзной промышленной системы, предназначенной для многократного выполнения одинаковых задач. В целом же можно повторить основной тезис: данные языки в существенной степени взаимозаменяемы.

2.3. Система контроля версий Git

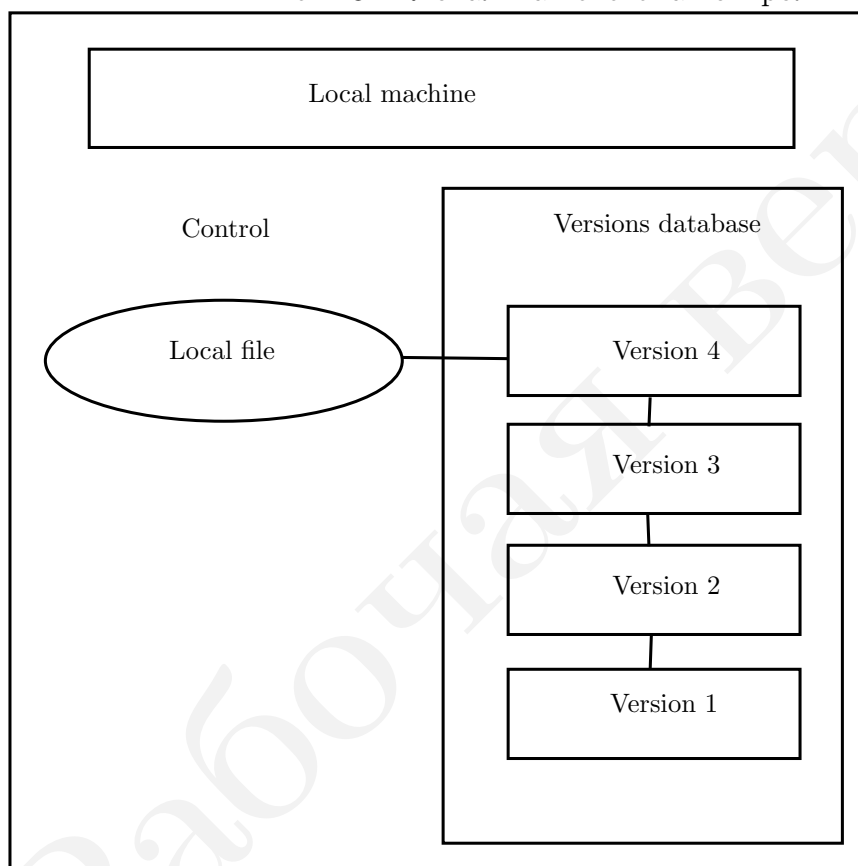
2.3.1. Общие сведения

Данный раздел не имеет отношения непосредственно к анализу данных, однако содержит сведения, полезные для комфортной работы при его осуществлении. Кроме того, использование систем контроля версий де факто является стандартом при любой серьёзной разработке, особенно в случае совместной работы над одним проектом нескольких аналитиков. Основная часть материала является пересказом [видеоурока](#) по работе с Git [133].

Система [Git](#) [26] — это одна из систем контроля версий. Система контроля версий — это система, записывающая изменения в файл или набор файлов в течение времени и позволяющая вернуться позже к определённой версии. Как правило подразумевается контроль версий файлов, содержащих исходный код программного обеспечения, хотя возможен контроль версий практически любых типов файлов [4]. Такие системы позволяют не только хранить версии файлов, но и содержат всю историю их изменения, позволяя отслеживать пошаговое изменение каждого бита файла.

1009 Это бывает особенно полезно в тех случаях, когда необходимо иметь возможность
1010 «откатить» изменения в случае наличия в них ошибок либо тогда, когда над одним
1011 и тем же проектом работает несколько разработчиков либо их команд. Конечно же
1012 можно просто создавать полные копии всех файлов проекта. Однако данный спо-
1013 соб полезен лишь для создания бэкапов на случай каких-то аварийных ситуаций.
1014 В обычной работе он, как минимум, неудобен, а, как максимум, просто не спосо-
1015 бен обеспечить пошаговое отслеживание изменений файлов и тем более слияние
1016 результатов нескольких команд, параллельно работающих над одними и теми же
1017 файлами. Для решения данной проблемы были разработаны локальные системы
1018 контроля версий, содержащие базу данных всех изменений в файлах, примерная
1019 схема организации которых показана на рисунке 2.3.1.

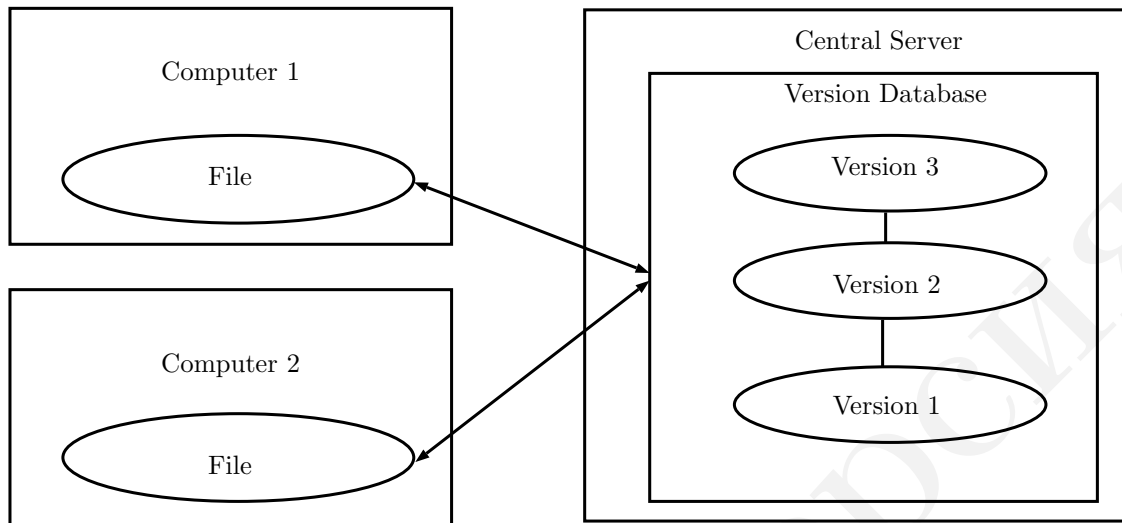
Рис. 2.3.1. Локальная система контроля версий



1020 Современные системы контроля версия бывают централизованными и распреде-
1021 лёнными. Первые устроены таким образом, что вся история изменений файлов хра-
1022 нится на центральном сервере, на который пользователи отправляют свои измене-
1023 ния, и с которого они их получают. Общая схема работы централизованной системы
1024 контроля версий приведена на рисунке 2.3.2 на следующей странице. Недостатком
1025 такой системы являет её зависимость от работы центрального сервера. В случае
1026 его остановки пользователи не смогут обрабатывать изменения, принимать и от-

1027 правлять их. Также существует риск полной потери всей истории в случае оконча-
1028 тельного отказа сервера.

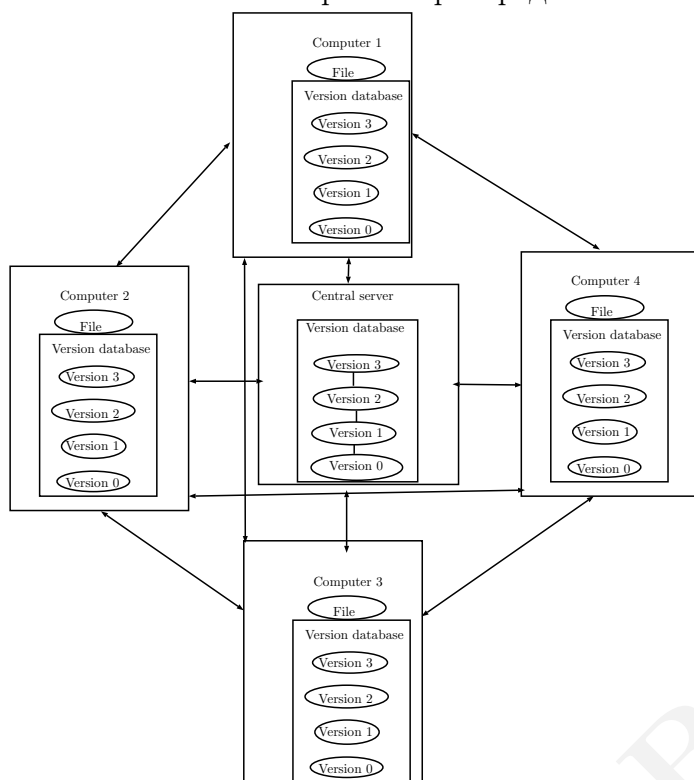
Рис. 2.3.2. Схема работы централизованной системы контроля версий



1029 Распределённые системы контроля версий лишены данного недостатка, поскольку у каждого пользователя хранится полная история изменений. В связи с этим
1030 каждый пользователь может продолжать работать с системой контроля при отсут-
1031 ствии связи с сервером. После восстановления работоспособности последнего,
1032 пользователь сможет синхронизировать свою историю изменений с другими разра-
1033 ботчиками. Даже в случае полного отказа сервера команда сможет просто перевести
1034 хранение на другой и продолжить работу в прежнем режиме. Общая схема работы
1035 распределённой системы приведена на рисунке 2.3.3.

1037 Особенностью работы системы Git является заложенный в ней принцип работы.
1038 В отличие от некоторых других систем контроля версий, принцип которых основан
1039 на хранении исходного файла и списка изменений к нему, Git хранит состояние
1040 каждого файла после его сохранения, создавая его «снимок». В терминологии Git
1041 каждый такой снимок называется commit. При этом создаются ссылки на каждый
1042 из файлов. В случае, если при создании нового commit Git обнаруживает, что какие-
1043 то файлы не были изменены, система не включает сами файлы в новый commit,
1044 а лишь указывает ссылку на последнее актуальное состояние файла из предыду-
1045 щего commit, обеспечивая таким образом эффективность дискового пространства.
1046 При этом каждый commit в целом ссылается на предыдущий, являющийся для него
1047 родительским. На рисунке 2.3.4 на с. 37 показана общая схема работы системы Git.
1048 Линиями со сплошным заполнением показана передача нового состояния файла, воз-
1049 никшего в результате внесения в него изменений, прерывистым — передача ссылки
1050 на состояние файла, не подвергавшегося изменениям, из прежнего commit. На мо-
1051 мент времени 0 (initial commit) все файлы находились в состоянии 0. Затем в файлы
1052 В и С были внесены изменения, тогда как файл А остался в прежнем состоянии.
1053 В процессе создания commit № 1 Git сделал снимок состояния файлов В1 и С1, а так-

Рис. 2.3.3. Схема работы распределённой системы контроля версий



же создал ссылку на состояние файла A0. Далее изменения были внесены в файл В. В процессе создания commit № 2 Git сохранил состояние файла B2, а также создал ссылки на состояния файлов A0 и C1 в предыдущем commit № 1. Затем были внесены изменения во все три файла, в результате чего на этапе создания commit № 3 Git сделал снимок состояний всех трёх файлов.

Внимательный читатель скорее всего обратит внимание на третий тип линий — пунктир, которому соответствует подпись «hash». Чтобы понять, каким образом в Git реализуется целостность версий, необходимо обратиться к понятию [хеш-функции](#) [9, 125].

2.3.2. Хеш-функции

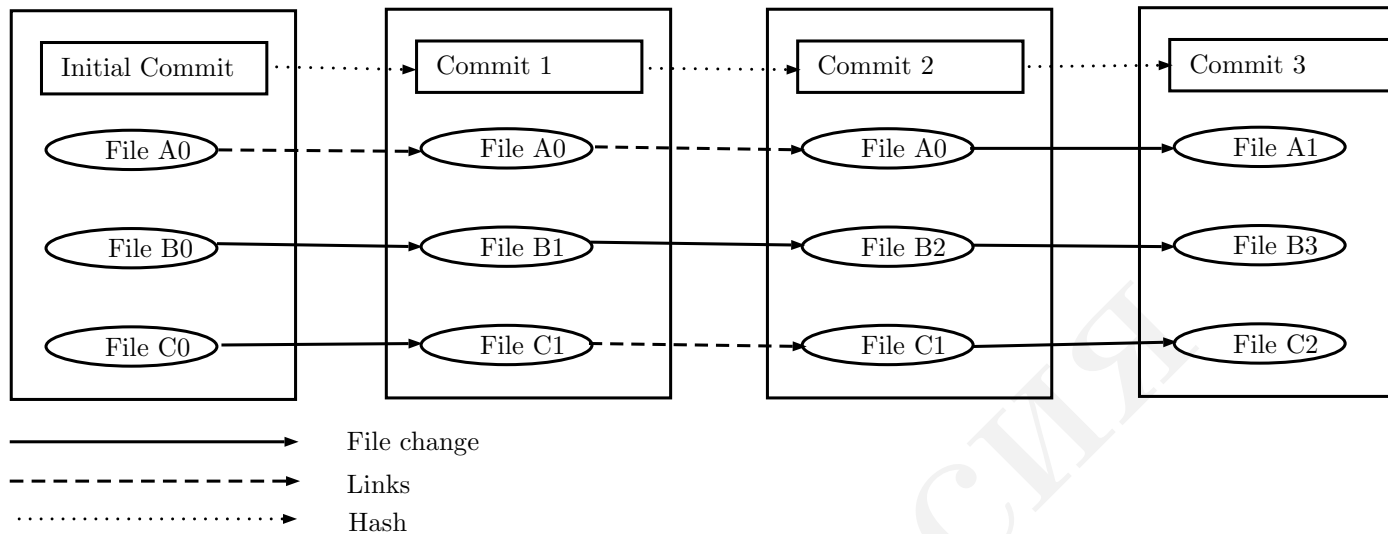
Приведём основные определения.

Хеш функция (функция свёртки) — функция, представляющая собой [детерминированный математический алгоритм](#) [107], осуществляющая преобразование данных произвольной длины в результирующую битовую строку фиксированной длины.

Хеширование — преобразование, осуществляемое хеш-функцией.

Сообщение (ключ, входной массив) — исходные данные.

Рис. 2.3.4. Общая схема работы Git



Хеш (хеш-сумма, хеш-код, сводка сообщения) — результат хеширования.

Согласно [Принципу Дирихле](#) [117], между хешем и сообщением в общем отсутствует однозначное соответствие. При этом, число возможных значений хеша меньше числа возможных значений сообщения. Ситуация, при которой применение одной и той же хеш-функции к двум различным сообщениям приводит к одинаковому значению хеша, называется «[коллизией хеш функции](#)» [110]. Т.е. коллизия имеет место тогда, когда $H(x) = H(y)$.

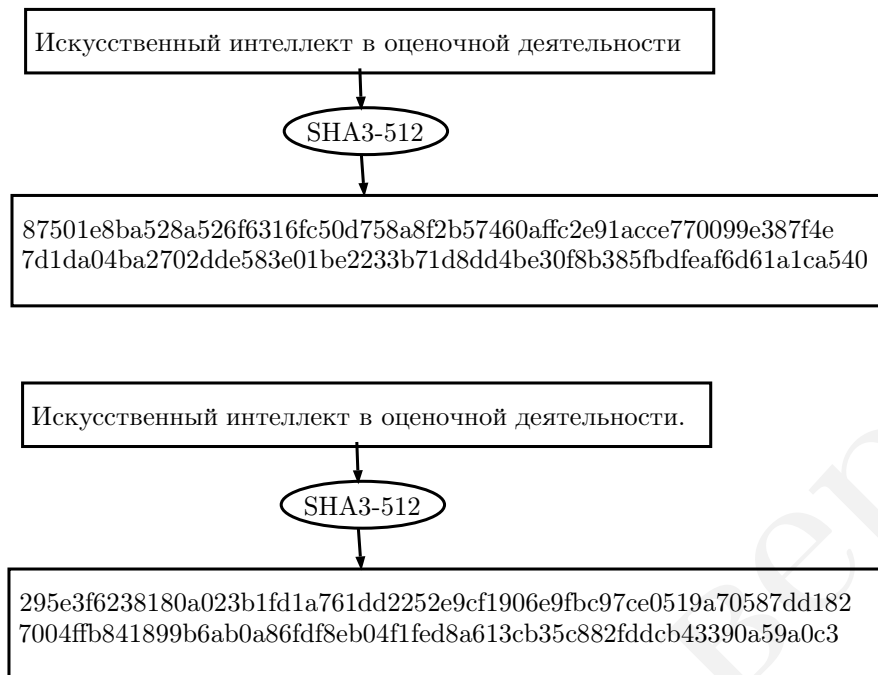
Теоретическая «идеальная» хеш-функция отвечает следующим требованиям:

- является детерминированной, то есть её применение к одному и тому же сообщению приводит к одному и тому же значению хеша любое число раз;
- значение хеша быстро вычисляется для любого сообщения;
- зная значение хеша, невозможно определить значение сообщения;
- невозможно найти такие два разных сообщения, применение хеширования к которым приводило бы к одинаковому значению хеша (т.е. идеальная хеш-функция исключает возможность возникновения коллизии);
- любое изменение сообщения (вплоть до изменения значения одного бита) изменяет хеш настолько сильно, что новое и старое значения выглядят никак не связанными друг с другом.

Как правило, название хеш-функции содержит значение длины результирующей битовой строки. Например хеш-функция [SHA3-512](#) [100] возвращает строку длиной в 512 бит. Воспользуемся [одним](#) [73] из онлайн-сервисов вычисления хеша и посчитаем его значение для названия данной книги. Как видно на рисунке [2.3.5 на следующей странице](#), результатом вычисления хеш-функции является строка длиной

1094 в 512 бит, содержащая 128 шестнадцатеричных чисел. При этом, можно наблюдать,
1095 что добавление точки в конце предложения полностью меняет значение хеша.

Рис. 2.3.5. Пример вычисления хеша



1096 Длина хеша в битах определяет максимальное количество сообщений, для кото-
1097 рых может быть вычислен уникальный хеш. Расчёт осуществляется по формуле.

$$2^n \quad (2.3.1)$$

1098 , где n — длина строки в битах.

1099 Так, для функции SHA3-512 число сообщений, имеющих уникальный хеш состав-
1100 ляет: $2^{512} \sim 1.340781 \times 10^{154}$. Таким образом, можно говорить о том, что современные
1101 хеш-функции способны генерировать уникальный хеш для сообщений любой дли-
1102 ны.

1103 Таким образом, Git в процессе создания нового commit сначала вычисляет его хеш-
1104 сумму, а затем фиксирует состояние. При этом в каждом commit присутствует ссыл-
1105 ка на предыдущий, также имеющий свою хеш-сумму. Таким образом, обеспечивается
1106 целостность истории изменений, поскольку значение хеш-суммы каждого после-
1107 дующего commit вычисляется на основе сообщения, содержащего в т. ч. свою хеш-
1108 сумму. В этом случае любая модификация содержимого данных, образующих лю-
1109 бой commit, неизбежно приведёт к изменению всех последующих хешей, что не оста-
1110 нется незамеченным.

2.3.3. Начало работы с Git и основные команды

Для того, чтобы начать работать с Git прежде всего его конечно же следует установить. Как правило, с этим не возникает никаких сложностей. Однако всё же вопросы установки Git кратко рассмотрены в подразделе [2.4.1 Git 74–75](#).

В данном подразделе преимущественно рассматриваются аспекты работы с ним через командную строку. Данный выбор обусловлен тем обстоятельством, что существует множество графических интерфейсов для работы с Git, которые активно развиваются, меняют дизайн и расширяют функционал. Кроме того, появляются новые продукты. Среди такого разнообразия всегда можно выбрать какой-то наиболее близкий для себя вариант. Таким образом, автор не видит смысла останавливаться на разборе какого-то конкретного графического интерфейса. Более важной задачей является изложение сути и основных принципов работы, понимание которых обеспечит успешную работу с Git безотносительно конкретных программных средств. Кроме того, следует отметить, что практически все современные IDE [109] имеют свои средства и интерфейс для работы с Git. В дальнейшем в главах, посвящённых непосредственно применению R и Python, будут рассмотрены вопросы использования Git средствами RStudio, Spyder и PyCharm.

В данном подразделе описывается работа с Git через командную строку в операционной системе Kubuntu. Большая часть изложенного применима для любой операционной системы. Для начала работы с Git откроем терминал и выполним три основные настройки, а именно укажем:

- имя пользователя;
- адрес электронной почты;
- текстовый редактор по умолчанию.

Для конфигурации Git существует специальная утилита *git config*, имеющая три уровня глобальности настроек:

```
$ git config --system
```

— системный уровень: затрагивает все репозитории всех пользователей системы;

```
$ git config --global
```

— глобальный уровень: затрагивает все репозитории конкретного пользователя системы;

```
$ git config --local
```

— локальный уровень: затрагивает конкретный репозиторий;

Представим, что необходимо задать общие настройки конкретного пользователя, т.е. использовать уровень `global`, что, может быть актуально, например, при использовании рабочего компьютера. Сделаем следующие настройки:

```
$ git config --global user.name "First.Second"
$ git config --global user.email user-adress@host.com
$ git config --global core.editor "kate"
```

— мы задали имя пользователя, адрес его электронной почты, отображаемые при выполнении `commit`, а также указали текстовый редактор по умолчанию. В данном случае был указан редактор `Kate`. Естественно можно указать любой другой удобный редактор. В случае использования операционной системы `Windows` необходимо указывать полный путь до исполняемого файла (имеет расширение `.exe`) текстового редактора, а также `a`. Например, в случае использования 64-х разрядной `Windows` и редактора `Notepad++` [65] команда может выглядеть так:

```
$ git config --global core.editor "'C:\Program Files\Notepad
\notepad.exe' -multiInst -notabbar -nosession -noPlugin"
```

— перечень команд для различных операционных систем и текстовых редакторов содержится на [соответствующей странице](#) сайта `Git` [26].

Для начала создадим тестовый каталог, с которым и будем работать в дальнейшем при обучении работе с `Git`. Зайдём в папку, в которой хотим создать каталог и запустим терминал в ней. После чего введём команду:

```
$ mkdir git-lesson
```

— мы только что создали новый каталог средствами командной строки. Затем введём команду:

```
$ cd git-lesson
```

— переходим в только что созданный каталог.

Для просмотра содержимого каталога используем следующую команду:

```
$ ls -la
```

— собственно самой командой является `ls`, а «`-la`» представляет собой её аргументы: «`-l`» — отвечает за отображение файлов и подкаталогов списком, а «`-a`» — за отображение скрытых файлов и подкаталогов.

Для создания репозитория введём команду:

```
$ git init
```

— `Git` ассоциирует текущую папку с новым репозиторием.

В случае, если всё прошло хорошо, терминал возвратит следующее сообщение:

```
> Initialized empty Git repository in /home/.../git-lesson/.
git/
```


Теперь ещё раз введём:

```
$ ls -la
```

— следует обратить внимание на то, что появилась папка `.git`, в которой и будет храниться вся история версий проекта, содержащегося в папке `git-lesson`.

Создадим первый файл внутри папки:

```
$ touch file1.py
```

— расширение указывает на то, что это файл языка Python.

Система Git уже должна была отследить наличие изменения состояния проекта, произошедшее вследствие создания нового файла. Для проверки изменений состояния используем команду:

```
$ git log
```

— и получим сообщение следующего содержания:

```
> fatal: your current branch 'master' does not have any
commits yet
```

— дело в том, что в истории изменений по-прежнему нет никаких записей.

Для получения дополнительных сведений используем команду:

```
$ git status
```

— терминал возвратит следующее сообщение:

```
> On branch master

No commits yet

Untracked files:
  (use "git add <file>..." to include in what will be
   committed)
   file1.py

nothing added to commit but untracked files present (use "
git add" to track)
```

— как видно, Git сообщает о том, что файл `file1.py` не отслеживается, кроме того, как следует из последней части сообщения терминала, в настоящее время вообще не фиксируются никакие изменения, поскольку ничего не было добавлено в лист отслеживания. При этом сам Git предлагает использовать команду `git add` для добавления файлов в него. Прежде чем сделать это, необходимо разобраться в том, в каких состояниях, с точки зрения Git, могут в принципе находиться файлы.

Все файлы, находящиеся в рабочем каталоге, могут иметь один из следующих статусов:

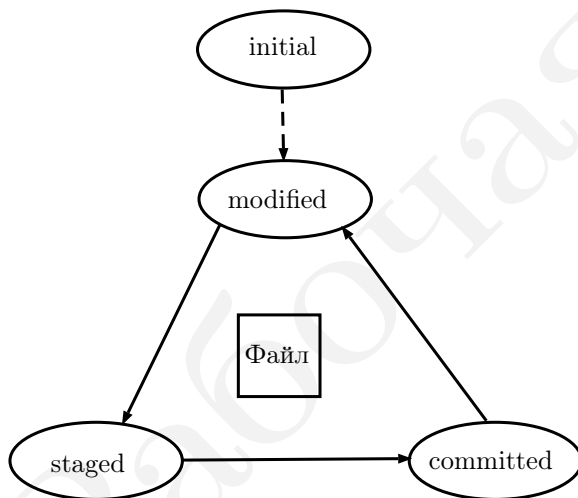
- `tracked` — отслеживаемые, т. е. находящиеся под версионным контролем;
- `untracked` — не отслеживаемые, т. е. не находящиеся под версионным контролем.

Ко второй категории, как правило, относятся временные файлы, например логи, хранение которых в репозитории нецелесообразно. Файлы первой категории могут находиться в одной из следующих состояний:

- `initial` — начальное состояние файла, в котором он находился в момент включения его в лист отслеживания, т. е. сообщения ему статуса `tracked`.
- `modified` — состояние файла после внесения в него изменений и его сохранения;
- `staged` — промежуточное состояние файла, в котором он находится после передачи его состояния Git, но до формирования последним его снимка.
- `committed` — состояние файла, зафиксированное Git, и представляющее его версию, к которой впоследствии будет возможно вернуться.

Соответственно после внесения новых изменений файл, находящийся в состоянии `committed`, переходит в состояние `modified`, после чего возможен новый цикл преобразований его статуса. Схема изменений состояния файлов приведена на рисунке 2.3.6.

Рис. 2.3.6. Схема состояний файлов в системе Git



Для перевода файла из состояния `modified` в состояние `staged` следует использовать команду

```
$ git add <file.name1> <file.name2>
```

— данная процедура также называется добавлением файла в индекс. Индекс — область памяти, в которой находятся файлы, подготовленные для включения в `commit`.

Далее для выполнения процедуры `commit` даётся команда

```
1275 $ git commit -m "message"
```

1278 — аргумент `-m` и следующее за ним сообщение служат для задания краткого описания того, какие изменения были внесены. Рекомендуется давать содержательные комментарии, позволяющие понять смысл изменений.

1281 Как видно, не обязательно совершать процедуру `commit` сразу в отношении всех файлов, находящихся в состоянии `modified`. Существует возможность группировать их и, посредством перевода конкретных файлов в состояние `staged`, формировать группы файлов, чьё состояние подлежит фиксации.

1285 Добавим файл `file.py` в индекс.

```
1286 $ git add file1.py
```

1289 Далее снова проверим статус:

```
1290 $ git status
```

1293 — на этот раз терминал возвратит новое сообщение:

```
1294 > On branch master
1295
1296 No commits yet
1297
1298 Changes to be committed:
1299   (use "git rm --cached <file>..." to unstage)
1300       new file:   file1.py
```

1303 Как можно видеть, теперь Git «видит» файл `file1.py` и готов сделать «снимок» нового состояния репозитория. Для выполнения процедуры `commit` введём команду:

```
1306 $ git commit -m "First commit"
```

1308 — мы только что сделали первый `commit`, т.е. зафиксировали состояние репозитория. Терминал возвратит следующее сообщение:

```
1310 > [master (root-commit) 1306b16] First commit
1311 1 file changed, 0 insertions(+), 0 deletions(-)
1312 create mode 100644 file1.py
```

1315 Теперь повторим ранее уже использованную команду:

```
1316 $ git log
```

1319 — терминал в отличие от первого раза, когда мы наблюдали сообщение о невозможности вывода сведений о событиях в репозитории, на этот раз возвращает осмысленное сообщение:

```
1322 > commit 1306b16f5fe40ccf8b141d716d9313df8e1983a1 (HEAD ->
1323     master) Author: Kirill Murashev <kirill.murashev@gmail.
1324     com>
```

```
Date:    Tue Aug 31 19:03:49 2021 +0200
First commit
```

— можно увидеть хеш-сумму данного commit, его автора, а также время создания commit и сопроводительное сообщение к нему. Для получения более детальных сведений можно использовать команду `git show`, сообщив ей в качестве аргумента хеш-сумму интересующего commit. Сделаем это, скопировав и вставив значение хеш-суммы:⁴

```
$ git show 1306b16f5fe40ccf8b141d716d9313df8e1983a1
```

— в качестве аргумента команды в данном случае была использована хеш-сумма. Терминал возвратит сообщение с данными об интересующем commit:

```
> commit 1306b16f5fe40ccf8b141d716d9313df8e1983a1 (HEAD ->
    master)
Author: Kirill Murashev <kirill.murashev@gmail.com>
Date:    Tue Aug 31 19:03:49 2021 +0300

    First commit

diff --git a/file1.py b/file1.py
new file mode 100644
index 0000000..e69de29
```

В дополнение к уже имеющимся данным приводятся сведения о том, какие именно изменения имели место. В данном случае видно, что имело место добавление в репозиторий нового файла.

Примерно такие же сведения можно получить в случае использования команды `git log` с аргументом `-p`.

```
$ git log -p

> commit 1306b16f5fe40ccf8b141d716d9313df8e1983a1 (HEAD ->
    master) Author: Kirill Murashev <kirill.murashev@gmail.
    com> Date:    Tue Aug 31 19:03:49 2021 +0300

    First commit

diff --git a/file1.py b/file1.py
new file mode 100644
index 0000000..e69de29
```

— в данном случае сообщения вообще идентичны.

⁴Для копирования и вставки в окне терминала следует использовать сочетания клавиш `ctrl+shift+c`, `ctrl+shift+v` соответственно.

Рассмотрим ещё одну полезную команду `git restore`. Данная команда возвращает состояние файла к тому состоянию, которое было зафиксировано при создании последнего commit. Рассмотрим пример. Откроем файл `file1.py` в редакторе Kate⁵ непосредственно из терминала:

```
$ kate file.py
```

— далее напишем в нём любой текст и сохраним файл. После чего проверим его статус с помощью уже известной команды `git status`:

```
$ git status

> On branch master
Changes not staged for commit:
  (use "git add <file>..." to update what will be committed
   )
  (use "git restore <file>..." to discard changes in working
   directory)
       modified:   file1.py

no changes added to commit (use "git add" and/or "git commit
-a")
```

— как видим, Git обнаружил изменение файла. Теперь введём команду:

```
$ git restore file.py
```

— файл, возвращён в состояние, в котором он находился на момент создания последнего commit, т. е. снова является пустым, в чём легко убедиться, открыв его.

Следующей рассматриваемой командой будет `git diff`. Данная команда позволяет понять, какие именно изменения были внесены в файл. Вновь откроем файл `file1.py` в текстовом редакторе. Введём в него текст, например «Liberte, egalite, fraternite». После чего сохраним файл. Выполним команду `git diff` и посмотрим на результат.

```
$git diff

> diff --git a/file1.py b/file1.py
index e69de29..72d6a2a 100644
--- a/file1.py
+++ b/file1.py
@@ -0,0 +1 @@
+Liberte, egalite, fraternite
```

— в нижней части сообщения терминала после символа «+» мы видим добавленный в файл текст. Git всегда отображает добавленный текст после знака «+», а удалённый после знака «-». Проверим статус файла:

⁵Естественно редактор может быть любой

```

1417 $ git status
1418
1419
1420 > On branch master
1421 Changes not staged for commit:
1422   (use "git add <file>..." to update what will be committed)
1423   (use "git restore <file>..." to discard changes in working
1424     directory)
1425         modified:   file1.py
1426
1427 no changes added to commit (use "git add" and/or "git commit
1428   -a")
1429

```

1430 — Git зафиксировал изменения файла. Теперь добавим файл в индекс, т. е. изменим его состояние на staged:

```

1431
1432 $ git add file1.py
1433

```

1435 — далее ещё раз проверим статус файла:

```

1436 $ git status
1437
1438
1439 > On branch master
1440 Changes to be committed:
1441   (use "git restore --staged <file>..." to unstage)
1442         modified:   file1.py
1443

```

1444 — Git перевёл файл в состояние staged. Для того, чтобы ещё раз посмотреть изменения в файле, находящемся в состоянии staged можно использовать ту же команду `git diff`, при условии сообщения ей аргумента `--staged`, без которого она не сможет отобразить изменения, поскольку они уже были включены в индекс.

```

1448 $ git diff --staged
1449
1450
1451 > diff --git a/file1.py b/file1.py
1452 index e69de29..d77d790 100644
1453 --- a/file1.py
1454 +++ b/file1.py
1455 @@ -0,0 +1 @@
1456 +Liberte, egalite, fraternite
1457

```

1458 Выполним commit:

```

1459 $ git commit -m "Second commit"
1460

```

1462 — терминал возвратит сообщение:

```

1463 > [master 700a993] Second commit
1464 1 file changed, 1 insertion(+)
1465

```

1467 — посмотрим на историю изменений:

```
1468 $ git log
1469
1470
1471 > commit 700a993db7c5f682c33a087cb882728adc485198 (HEAD ->
1472     master)
1473 Author: Kirill Murashev <kirill.murashev@gmail.com>
1474 Date:    Tue Aug 31 20:51:06 2021 +0200
1475
1476     Second commit
1477
1478 commit 1306b16f5fe40ccf8b141d716d9313df8e1983a1
1479 Author: Kirill Murashev <kirill.murashev@gmail.com>
1480 Date:    Tue Aug 31 19:03:49 2021 +0200
1481
1482     First commit
1483
```

1484 — можно наблюдать сведения о двух выполненных commit.

1485 В случае использования той же команды с аргументом `-p` можно увидеть всю
1486 историю конкретных изменений.

```
1487 $ git log -p
1488 > commit 700a993db7c5f682c33a087cb882728adc485198 (HEAD ->
1489     master)
1490 Author: Kirill Murashev <kirill.murashev@gmail.com>
1491 Date:    Tue Aug 31 20:51:06 2021 +0300
1492
1493     Second commit
1494
1495 diff --git a/file1.py b/file1.py
1496 index e69de29..d77d790 100644
1497 --- a/file1.py
1498 +++ b/file1.py
1499 @@ -0,0 +1 @@
1500 +Liberte, egalite, fraternite
1501
1502 commit 1306b16f5fe40ccf8b141d716d9313df8e1983a1
1503 Author: Kirill Murashev <kirill.murashev@gmail.com>
1504 Date:    Tue Aug 31 19:03:49 2021 +0300
1505     First commit
1506 diff --git a/file1.py b/file1.py
1507 new file mode 100644
1508 index 0000000..e69de29
1509
1510
```

1511 Существует упрощённый способ передачи Git сведений для совершения commit.
1512 Вместо последовательного ввода команд `git add` с указанием перечня файлов и `git`

commit можно использовать единую команду `git commit` с аргументами `-am`. Вторым аргументом, как уже было сказано ранее, необходим для формирования сообщения, сопровождающего commit. Первый же заменяет собой предварительное использование команды `git add`, указывая Git на необходимость включения в индекс всех отслеживаемых файлов, т. е. имеющих статус `tracked`. Внесём любые изменения в файл `file1.py`. Проверим наличие изменений:

```

1519 $ git status
1520
1521
1522 > On branch master
1523 Changes not staged for commit:
1524   (use "git add <file>..." to update what will be committed)
1525   use "git restore <file>..." to discard changes in working
1526       directory)
1527       modified:   file1.py
1528
1529 no changes added to commit (use "git add" and/or "git commit
1530 -a")
1531

```

— после чего выполним добавление в индекс и commit одной командой.

```

1532 $ git commit -am "Third commit"
1533
1534 > [master fbff919] Third commit
1535 1 file changed, 1 insertion(+)
1536
1537

```

— проверим историю:

```

1538 $ git log -p
1539
1540
1541
1542 > commit fbff919fab14ab6d41c993d3b86253c41037e075 (HEAD ->
1543     master)
1544 Author: Kirill Murashev <kirill.murashev@gmail.com>
1545 Date:   Tue Aug 31 21:25:45 2021 +0300
1546
1547     Third commit
1548
1549 diff --git a/file1.py b/file1.py
1550 index d77d790..bf6409f 100644
1551 --- a/file1.py
1552 +++ b/file1.py @@ -1,2 @@
1553     Liberte, egalite, fraternite
1554 +Жизнь, свобода, собственность
1555
1556 commit 700a993db7c5f682c33a087cb882728adc485198
1557 Author: Kirill Murashev <kirill.murashev@gmail.com>
1558 Date:   Tue Aug 31 20:51:06 2021 +0300

```

```

1559
1560     Second commit
1561
1562 diff --git a/file1.py b/file1.py
1563 index e69de29..d77d790 100644
1564 --- a/file1.py
1565 +++ b/file1.py
1566 @@ -0,0 +1 @@
1567 +Liberte, egalite, fraternite
1568
1569 commit 1306b16f5fe40ccf8b141d716d9313df8e1983a1
1570 Author: Kirill Murashev <kirill.murashev@gmail.com>
1571 Date:   Tue Aug 31 19:03:49 2021 +0300
1572
1573     First commit
1574
1575 diff --git a/file1.py b/file1.py
1576 new file mode 100644
1577 index 0000000..e69de29
1578

```

1579 — можно наблюдать уже три commit.

1580 Следующей полезной командой является `git mv`. Данная команда позволяет, в част-
 1581 ности, переименовывать либо перемещать файлы. При этом её выполнение автома-
 1582 тически переводит файл в состояние `staged`, минуя состояние `modified`. Выполним
 1583 переименование:

```

1584 $ git mv file1.py file-1.py
1585
1586

```

1587 — затем проверим состояние:

```

1588 $ git status
1589
1590
1591 > On branch master
1592 Changes to be committed:
1593   (use "git restore --staged <file>..." to unstage)
1594     renamed:   file1.py -> file-1.py
1595

```

1596 — как можно увидеть, файл с новым именем готов к `commit`. Выполним `commit`.

```

1597 $ git commit -m "Fourth commit"
1598
1599
1600 > [master 284073c] Fourth commit
1601 1 file changed, 0 insertions(+), 0 deletions(-)
1602 rename file1.py => file-1.py (100%)
1603

```

1604 — изменения файла зафиксированы.

1605 Следующей заслуживающей внимания командой является `git rm`. Данная ко-
1606 манда удаляет файл.

```
1607 $ git rm file-1.py
```

1610 — проверим выполнение операции:

```
1611 $ git status
1612
1613 > On branch master
1614 Changes to be committed:
1615   (use "git restore --staged <file>..." to unstage)
1616     deleted:    file-1.py
```

1619 — как видно из сообщения Git в терминале, существует возможность восстановить
1620 удалённый файл в том состоянии, которое было зафиксировано при выполнении
1621 последнего commit. Выполним команду для восстановления файла:

```
1622 $ git restore --staged file-1.py
```

1625 — затем проверим его состояние:

```
1626 $ git status
1627
1628 > On branch master
1629 Changes not staged for commit:
1630   (use "git add/rm <file>..." to update what will be
1631    committed)
1632   (use "git restore <file>..." to discard changes in working
1633    directory)
1634     deleted:    file-1.py
1635
1636 no changes added to commit (use "git add" and/or "git commit
1637   -a")
```

1640 — как следует из сообщения Git, файл `file-1.py` больше не находится в индексе,
1641 для его возвращения туда необходимо выполнить команду `git restore` без указа-
1642 ния каких-либо аргументов.

```
1643 $ git restore file-1.py
```

1646 — ещё раз проверим состояние:

```
1647 $ git status
1648
1649 > On branch master nothing to commit, working tree clean
```

1652 — файл снова включён в индекс, его состояние соответствует состоянию, зафикси-
1653 рованному при выполнении последнего commit. Сам файл при этом вновь присут-
1654 ствует в каталоге.

Команда `git rm` также может быть использована для передачи файлу статуса `untracked` без его удаления из каталога. Для этого ей необходимо сообщить аргумент `--cached`.

```
$ git rm --cached file-1.py
>rm 'file-1.py'
```

— файл был исключён из индекса, а также из списка отслеживания, но при этом остался в каталоге, в чём можно легко убедиться:

```
$ git status
> On branch master
Changes to be committed:
  (use "git restore --staged <file>..." to unstage)
    deleted:    file-1.py

Untracked files:
  (use "git add <file>..." to include in what will be
   committed)
    file-1.py
```

— есть изменения, доступные для `commit`, а также в каталоге присутствует неотслеживаемый файл (статус `untracked`).

```
$ ls -la
> total 0\
drwx----- 1 user.name root  0 jaan  1 1970 .
drwx----- 1 user.name root  0 jaan  1 1970 ..
-rwx----- 1 user.name root 84 sept  1 19:08 file-1.py
drwx----- 1 user.name root  0 jaan  1 1970 .git
```

— файл присутствует в каталоге.

Выполним `commit`:

```
$ git commit -m "Fifth commit"
> [master 7abee55] Fifth commit
1 file changed, 2 deletions(-)
delete mode 100644 file-1.py
```

— далее посмотрим историю изменений:

```
$ git log
> commit 7abee55d2631cf7cf2e94e58f30f36b2be807948 (HEAD ->
   master)
Author: Kirill Murashev <kirill.murashev@gmail.com>
Date:   Wed Sep 1 19:52:04 2021 +0300
```

```

1703     Fifth commit
1704
1705 commit 284073c521af8b73e16f324698f24040e4b9ee7e
1706 Author: Kirill Murashev <kirill.murashev@gmail.com>
1707 Date:   Wed Sep 1 18:16:46 2021 +0300
1708
1709     Fourth commit
1710
1711 commit fbff919fab14ab6d41c993d3b86253c41037e075
1712 Author: Kirill Murashev <kirill.murashev@gmail.com>
1713 Date:   Tue Aug 31 21:25:45 2021 +0300
1714
1715     Third commit
1716
1717 commit 700a993db7c5f682c33a087cb882728adc485198
1718 Author: Kirill Murashev <kirill.murashev@gmail.com>
1719 Date:   Tue Aug 31 20:51:06 2021 +0300
1720
1721     Second commit
1722
1723 commit 1306b16f5fe40ccf8b141d716d9313df8e1983a1
1724 Author: Kirill Murashev <kirill.murashev@gmail.com>
1725 Date:   Tue Aug 31 19:03:49 2021 +0300
1726
1727     First commit
1728

```

1729 — проверим наличие файла в каталоге:

```

1730
1731 $ ls -la
1732 > total 0
1733 drwx----- 1 user.name root 0 jaan 1 1970 .
1734 drwx----- 1 user.name root 0 jaan 1 1970 ..
1735 -rwx----- 1 user.name root 84 sept 1 19:08 file-1.py
1736 drwx----- 1 user.name root 0 jaan 1 1970 .git
1737

```

1738 — а также его статус:

```

1739
1740 $ git status
1741 > On branch master
1742 Untracked files:
1743   (use "git add <file>..." to include in what will be
1744     committed)
1745     file-1.py
1746
1747 nothing added to commit but untracked files present (use "
1748   git add" to track)

```

1749

1750 — файл присутствует в каталоге и имеет статус `untracked`.
1751 Вернём файл в индекс.

1752

1753

1754

```
$ git add file-1.py
```

1755 — файл вновь имеет статус `tracked`.

1756

2.3.4. Исключение файлов из списка отслеживания

1757 В процессе разработки нередко возникают файлы, отслеживание которых скорее
1758 всего является нецелесообразным, например файлы, содержащие логи. При этом
1759 их постоянное присутствие в списке файлов, имеющих статус `untracked`, осложняет
1760 работы и также является нежелательным. В связи с этим существует механизм
1761 исключения ряда файлов или подкаталогов из под всей системы версионирования,
1762 называемый *gitignore*.

1763 Выполним ряд процедур. До этого все действия выполнялись путём последова-
1764 тельного ввода команд. В данном случае будет показано, как можно использовать
1765 заготовленные скрипты. Использование скриптов является очень удобным тогда,
1766 когда существует необходимость многократного ввода длинной последовательности
1767 команд. В рассматриваемом примере будет рассмотрена последовательность всего
1768 из пяти команд. Для создания скрипта необходимо написать его текст в текстовом
1769 редакторе, сохранить файл с расширением `txt` (например `script1.txt`), после чего за-
1770 пустить терминал в каталоге с файлом и указать системе на то, что данный файл
1771 является исполняемым, т. е. передать ему права `execute`. Напишем скрипт:

1772

1773

1774

1775

1776

1777

1778

1779

1780

1781

1782

1783

1784

1785

1786

1787

1788

1789

1790

1791

1792

```
# создаём подкаталог  
mkdir log  
# переходим в новый подкаталог  
cd log/  
# создаём файл  
touch log.txt  
# возвращаемся в каталог верхнего уровня  
cd ..  
# проверяем статус  
git status
```

1784 — смысл того, что выполняет команда раскрыт в комментарии, предшествующем
1785 ей. Следует обратить внимание на то, что команды, передаваемые терминалу пи-
1786 шутся на языке [Bash \[82\]](#), в котором игнорируется всё, что написано в строке после
1787 символа «`#`». Передадим файлу права `execute` путём ввода команд в терминала, за-
1788 пущенном из каталога, содержащего файл. Можно использовать любую (двоичную
1789 либо символическую) запись:

1790

1791

1792

```
$ chmod u+x script
```

1793 — либо:

```
1794 $ chmod 744 script
```

1797 — для проверки наличия прав в системе Kubuntu и многих других можно использовать команду:

```
1799 $ ls -l script1
```

1802 — в случае наличия прав execute терминал возвратит ответ, содержащий имя файла, выделенное **зелёным** цветом.

1804 Теперь следует вернуться в окно терминала, запущенное в каталоге изучаемого репозитория после чего просто ввести нём полный путь до созданного скрипта:

```
1806 $ ~/.../Scripts/script1
```

1809 — в случае правильных действий терминал возвратит сообщение:

```
1810 > On branch maste
1811 r Changes to be committed:
1812   (use "git restore --staged <file>..." to unstage)
1813       new file:   file-1.py
1814
1815 Untracked files:
1816   (use "git add <file>..." to include in what will be
1817       committed)
1818       log/
1819
1820
```

1821 В данном случае автор использовал заготовленный bash скрипт. Аналогичного результата можно добиться путём простого последовательного ввода команд. Подробнее о запуске скриптов в операционных системах, основанных на ядре Linux, можно прочитать, например [здесь](#) [130]. Возвращаясь к теме Git, отметим, что в каталоге появилась неотслеживаемая папка log. Создадим файл с именем .gitignore:

```
1826 $ kate .gitignore
```

1829 — при этом сразу же откроется окно текстового редактора. Следует сделать небольшое отступление и сказать о том, что состав файлов и папок, подлежащих исключению из списка, подлежащего версионированию, в существенной степени зависит от используемого языка программирования. В дальнейшем будут рассмотрены вопросы автоматизации создания файла .gitignore. Сейчас же кратко рассмотрим заготовленные файлы для языков Python и R. Ниже приводится примерное содержание файла .gitignore, предназначенного для репозитория, содержащего код на языке Python:

```
1837 # Byte-compiled / optimized / DLL files
1838 __pycache__ /
1839 *.py[cod]
1840 *$py.class
```



```
1842
1843 # C extensions
1844 *.so
1845
1846 # Distribution / packaging
1847 .Python
1848 build/
1849 develop-eggs/
1850 dist/
1851 downloads/
1852 eggs/
1853 .eggs/
1854 lib/
1855 lib64/
1856 parts/
1857 sdist/
1858 var/
1859 wheels/
1860 share/
1861 python-wheels/
1862 *.egg-info/
1863 .installed.cfg
1864 *.egg MANIFEST
1865
1866 # PyInstaller
1867 # Usually these files are written by a python script from a
1868     template
1869 # before PyInstaller builds the exe, so as to inject date/
1870     other infos into it.
1871 *.manifest
1872 *.spec
1873
1874 # Installer logs
1875 pip-log.txt
1876 pip-delete-this-directory.txt
1877
1878 # Unit test / coverage reports
1879 htmlcov/
1880 .tox/
1881 .nox/
1882 .coverage
1883 .coverage.*
1884 .cache nosetests.xml
1885 coverage.xml
```

```
147 *.cover
148 *.py,cover
149 .hypothesis/
150 .pytest_cache/
151 cover/
152
153 # Translations
154 *.mo
155 *.pot
156
157 # Django stuff:
158 *.log
159 local_settings.py
160 db.sqlite3
161 db.sqlite3-journal
162
163 # Flask stuff:
164 instance/
165 .webassets-cache
166
167 # Scrappy stuff:
168 .scrappy
169
170 # Sphinx documentation
171 docs/_build/
172
173 # PyBuilder
174 .pybuilder/
175 target/
176
177 # Jupyter Notebook
178 .ipynb_checkpoints
179
180 # IPython
181 profile_default/
182 ipython_config.py
183
184 # pyenv
185 # For a library or package, you might want to ignore these
1925 files since the code is
186 # intended to run in multiple environments; otherwise,
1927 check them in:
187 # .python-version
188 # pipenv
```

```
1930 # According to pypa/pipenv#598, it is recommended to
1931 include Pipfile.lock in version control.
1932 # However, in case of collaboration, if having platform-
1933 specific dependencies or dependencies
1934 # having no cross-platform support, pipenv may install
1935 dependencies that don't work, or not
1936 # install all needed dependencies.
1937 #Pipfile.lock
1938 # PEP 582; used by e.g. github.com/David-OConnor/pyflow
1939 __pypackages__/_
1940
1941 # Celery stuff
1942 celerybeat-schedule
1943 celerybeat.pid
1944
1945 # SageMath parsed files
1946 *.sage.py
1947
1948 # Environments
1949 .env
1950 .venv
1951 env/
1952 venv/
1953 ENV/
1954 env.bak/
1955 venv.bak/
1956
1957 # Spyder project settings
1958 .spyderproject
1959 .spyproject
1960
1961 # Rope project settings
1962 .ropeproject
1963
1964 # mkdocs documentation
1965 /site
1966
1967 # mypy
1968 .mypy_cache/
1969 .dmypy.json dmypy.json
1970
1971 # Pyre type checker
1972 .pyre/
1973
```

```

130 # pytype static type analyzer
131 .pytype/
132
133 # Cython debug symbols
134 cython_debug/
135
136 # учебная строка, добавлена автором
137 log/

```

— можно сказать, что файл содержит в себе в т. ч. набор относительно простых регулярных выражений. В частности символ «*» означает возможность наличия любых символов. Заключение последовательности символов в квадратные скобки означает возможность присутствия на данном месте любого из них. В частности в строке 3 содержится указание на необходимость игнорирования файлов, имеющих любое имя и одно из следующих расширений: .рус, .руо, .руд.

Примерное содержание файла .gitignore, предназначенного для репозитория, содержащего код на языке R:

1991		
1992	# History files	1
1993	.Rhistory	2
1994	.Rapp.history	3
1995		4
1996	# Session Data files	5
1997	.RData	6
1998		7
1999	# User-specific files	8
2000	.Ruserdata	9
2001		10
2002	# Example code in package build process	11
2003	*-Ex.R	12
2004		13
2005	# Output files from R CMD build	14
2006	/*.tar.gz	15
2007		16
2008	# Output files from R CMD check	17
2009	/*.Rcheck/	18
2010		19
2011	# RStudio files	20
2012	.Rproj.user/	21
2013		22
2014	# produced vignettes	23
2015	vignettes/*.html	24
2016	vignettes/*.pdf	25
2017		26
2018	# OAuth2 token, see https://github.com/hadley/http/releases/	27

2019	<code>tag/v0.3</code>	
2020	<code>.httr-oauth</code>	28
2021		29
2022	<code># knitr and R markdown default cache directories</code>	30
2023	<code>*_cache/</code>	31
2024	<code>/cache/</code>	32
2025		33
2026	<code># Temporary files created by R markdown</code>	34
2027	<code>*.utf8.md</code>	35
2028	<code>*.knit.md</code>	36
2029		37
2030	<code># R Environment Variables</code>	38
2031	<code>.Renviron</code>	39
2032		40
2033	<code># pkgdown</code>	41
2034	<code>site docs/</code>	42
2035		43
2036	<code># translation temp files</code>	44
2037	<code>po/*~</code>	45
2038		46
2039	<code>#учебная строка, добавлена автором</code>	47
2040	<code>log/</code>	48
2041		

2042 — используем любой из указанных файлов, сохраним его и проверим статус:

```

2043 $ git status
2044 > On branch master
2045 Changes to be committed:
2046   (use "git restore --staged <file>..." to unstage)
2047     new file:   file-1.py
2048
2049 Untracked files:
2050   (use "git add <file>..." to include in what will be
2051     committed)
2052     .gitignore
2053
2054 
```

2055 — как видим папка log пропала и появился файл .gitignore. Добавим его в индекс:

```

2056 $ git add .gitignore
2057
2058 
```

2059 — а затем выполним commit:

```

2060 $ git commit -m "Sixth commit"
2061 > [master e4adf82] Sixth commit
2062 2 files changed, 142 insertions(+)
2063 create mode 100644 .gitignore
2064 create mode 100644 file-1.py
2065
2066 
```

— теперь в случае создания в каталоге любого файла, чьё имя подпадает под правила, описанные в файле `.gitignore`, он сразу же исключается из списка наблюдения со стороны системы версионирования. Забегая вперёд, можно сказать, что, чаще всего отсутствует необходимость создавать такой файл вручную. Данная функция реализована во многих IDE и будет рассмотрена далее.

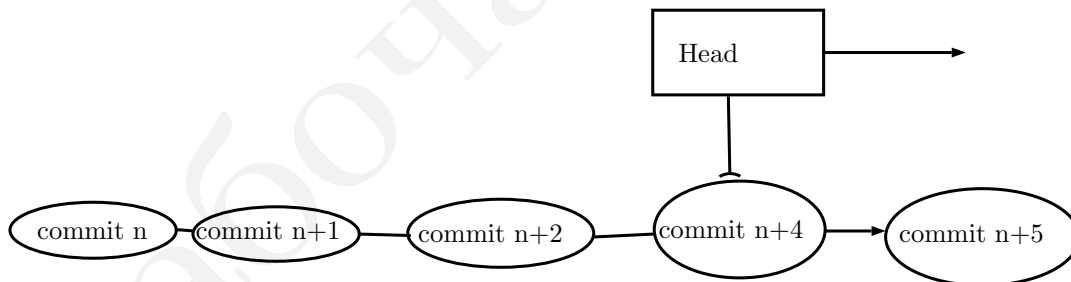
2.3.5. Ветки проекта, указатели `branch` и `Head`

В предыдущих подразделах рассматривалась линейная модель созданий версий, которые последовательно формировались одна за другой путём проведения процедуры `commit`. Git позволяет осуществлять ветвление версий. Посмотрим на текущий статус репозитория:

```
$ git status
> On branch master nothing to commit, working tree clean
```

— обратим внимание на сообщение, возвращённое терминалом, содержащее ссылку на некую `branch master`. Для того, чтобы разобраться в данном вопросе, следует вспомнить основные принципы работы Git, описанные в подразделах 2.3.1–2.3.2 на с. 33–38. Каждый `commit` имеет хеш-сумму, содержащую в т.ч. ссылку на предыдущий `commit`. Таким образом формируется неразрывная цепочка версий. Помимо этого в Git реализована работа указателя `Head`, представляющего собой метку, указывающую на один из `commit`. Местонахождение этой метки указывает Git, в каком именно состоянии репозиторий находится в данный момент. При каждом выполнении `commit` указатель `Head` смещается на новый `commit`. Схема работы указателя `Head` показана на рисунке 2.3.7.

Рис. 2.3.7. Схема работы указателя `Head`



Установить текущее местонахождение указателя `Head` можно с помощью команды `git log`.

```
$ git log
> commit e4adf8280c5d95a6f5796dba8e028012565de958
(HEAD -> master)
Author: Kirill Murashev <kirill.murashev@gmail.com>
```

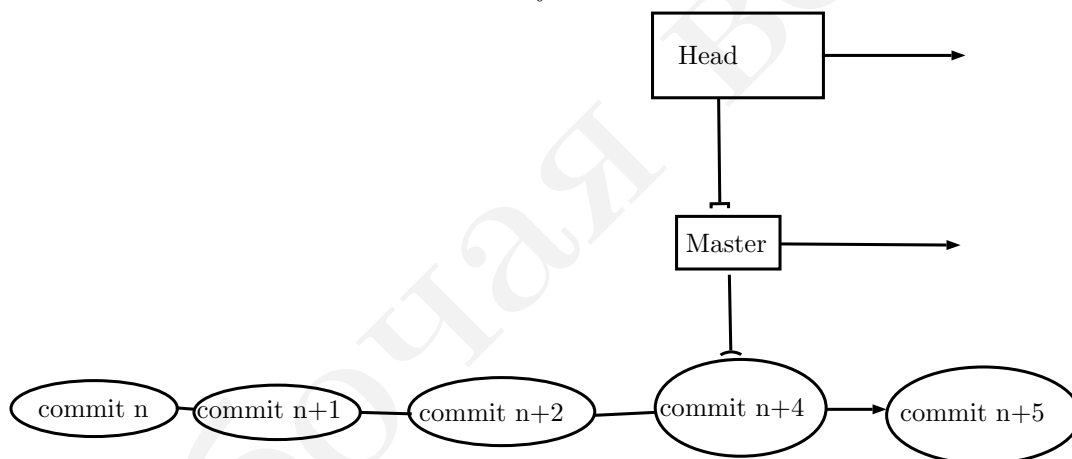
```

2099 Date:    Thu Sep 2 20:42:49 2021 +0300
2100
2101     Sixth commit
2102
2103 commit 7abee55d2631cf7cf2e94e58f30f36b2be807948 Author:
2104     Kirill Murashev <kirill.murashev@gmail.com>
2105 Date:    Wed Sep 1 19:52:04 2021 +0300
2106
2107     Fifth commit
2108
2109 commit 284073c521af8b73e16f324698f24040e4b9ee7e
2110
2111 :...skipping...

```

— как следует из ответа терминала, указатель Head находится на последнем шестом commit ветки Master. При этом ветка также представляет собой некий указатель. Таким образом, схема организации указателей выглядит так, как это показано на рисунке 2.3.8.

Рис. 2.3.8. Схема указателей Head и Branch



При наличии достаточной степени развития проекта хорошей практикой считается хранение стабильной версии в ветке Master (в современных системах часто используется наименование Main).⁶

При этом, для новых изменений, находящихся в стадии разработки и тестирования, рекомендуется использовать отдельную ветку. Для создания новой ветки следует использовать команду `git branch <name>`:

```

2123 $ git branch Develop
2124
2125

```

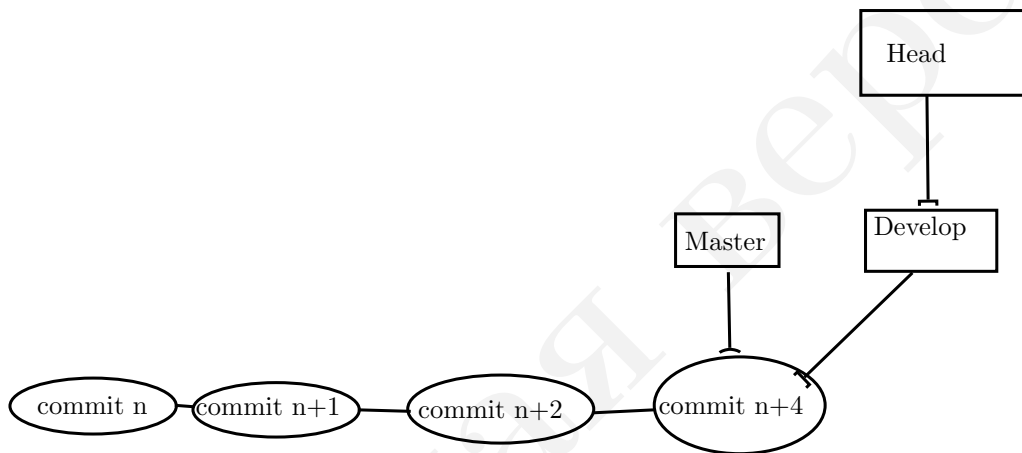
⁶ Данное решение обусловлено политическими причинами, поскольку слово Master может ассоциироваться с рабовладением.

— была создана новая ветка Develop. Для перемещения указателя Head на неё используем команду:

```
$ git checkout Develop
```

— состояние репозитория выглядит следующим образом: см. рисунок 2.3.9. Теперь все последующие commit будут сопровождаться указателем ветки Develop, тогда как указатель Master останется на прежнем месте. В случае обратного перемещения указателя Head на ветку Master состояние файлов проекта вернётся к тому, каким оно было в момент создания commit, на который теперь указывает Head. При этом все изменения, сделанные в ветке Develop будут сохранены в ней и доступны в случае перемещения Head на них. После определённого количества перемещений и доработок проект может выглядеть, например так, как показано на рисунке 2.3.10 на следующей странице.

Рис. 2.3.9. Состояние репозитория после переноса указателя Head на ветку Develop



Предположим, что, достигнув состояния, показанного на рисунке на рисунке 2.3.10 на следующей странице, оценщик приходит к выводу о необходимости слияния всех веток в ветку master. Сначала можно посмотреть, какие ветки в принципе существуют.

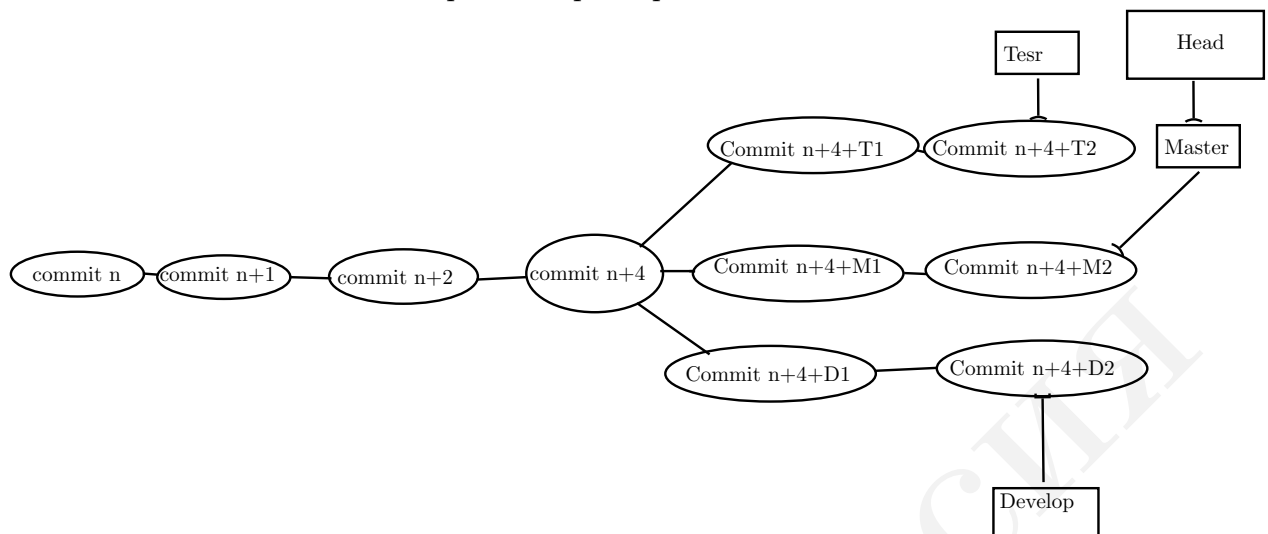
```
$ git branch -a
>
  develop
* master
  test
```

Как видно, указатель Head уже находится на целевой ветке. Если это не так, его следует туда перенести:

```
$ git checkout master
```

После этого выполняем команду `git merge`, в качестве аргумента которой используется имя ветки, которую предполагается объединить с master:

Рис. 2.3.10. Состояние репозитория при наличии нескольких веток



```
2158 $ git merge develop
2159
2160
```

2161 — в случае, когда последний commit из ветки master является прямым родителем
 2162 для ветки develop объединение происходит путём простого перемещения указателя
 2163 Head на master, а затем их совместного перемещения на последний commit ветки
 2164 develop. Данный способ объединения называется fast forward. После такого объеди-
 2165 нения может быть целесообразным удалить ветку develop, поскольку дальнейшая
 2166 работа будет вестись в master. Для этого следует выполнить команду:

```
2167 $ git branch -d develop
2168
```

2170 — теперь все изменения, ранее сделанные в ветке develop по-прежнему доступны
 2171 уже в ветке master. В случае, когда объединяемые ветки не являются родитель-
 2172 ской и дочерней относительно друг друга процедура объединения происходит более
 2173 сложным образом. При этом используется та же команда. В результате её выпол-
 2174 нения формируется новый commit, называемый merge commit. При этом проис-
 2175 ходит перемещение указателей master и Head на данный commit. В случае, если
 2176 commits, являющиеся родительскими по отношению к merge commit имеют толь-
 2177 ко непересекающиеся дополнения относительно последнего общего родительского
 2178 commit и не имеют взаимоисключающих правок, объединение происходит автома-
 2179 тически и не требует внимания пользователя. Ситуация, при которой имеет место
 2180 т. н. merge conflict, будет рассмотрена в подразделе 2.3.7 на с. 73–73. Последова-
 2181 тельное выполнение команд git branch и git checkout можно заменить одним
 2182 командой:

```
2183 $ git checkout -b <branch.name>
2184
2185
```

2186 — в случае необходимости отведения новой ветки не от того commit, на который

указывает Head вторым аргументом этой команды должна быть хеш-сумма того commit, от которого необходимо отвести ветку.

2.3.6. Работа с Github

2.3.6.1. Начало

В материале, изложенном выше в подразделах 2.3.3–2.3.5 2.3.3 на с. 39–64, речь шла о работе с локальным репозиторием, хранящимся на компьютере пользователя. При этом при командной работе часто требуется наличие общего доступа к рабочему каталогу. Также наличие удалённой версии репозитория позволяет распространять разработки на широкую аудиторию. Кроме того, наличие удалённого репозитория позволяет иметь дополнительный бэкап, не зависящий от физического устройства пользователя. Следуя **принципу KISS**, положенному в основу данной работы, в настоящем разделе будет рассмотрена работа с наиболее популярным сервисом удалённых репозитория — **GitHub** [29]. Следует отметить, что существует значительное количество альтернатив, кроме того существует возможность хранения удалённого репозитория на собственном удалённом сервере.

Для начала работы с GitHub необходимо осуществить регистрацию, которая вряд ли может вызвать у кого затруднение в 2021 году. Для создания своего первого репозитория необходимо в меню профиля выбрать пункт Your Repositories и далее создать свой, что также вряд ли может вызвать затруднения. В появившемся меню следует ввести имя репозитория латинскими символами, затем выбрать тип репозитория: публичный либо приватный. В первом случае доступ к репозиторию (но его изменению его содержимого) будет о неограниченного круга пользователей. Для доступа к материалам достаточно иметь ссылку на репозиторий. Во втором случае доступ даже к просмотру будут иметь только те, кому будет предоставлены соответствующие права. Следующие пункты меню позволяют добавить файл README, содержащий основные сведения о проекте, файл .gitignore, сформированный по шаблону, разработанному для конкретного языка, а также выбрать лицензию, на условиях которой возможно легальное использование продукта.

Для обеспечения связи между локальным репозиторием и его удалённой версией необходимо зайти в него и выбрать меню **Code**. В данном меню можно выбрать одно из трёх средств передачи данных:

- протокол **HTTPS** [90];
- протокол **SSH** [101];
- средства командной строки GitHub CLI, в свою очередь также реализующие передачу данных посредством протоколов:
 - **HTTPS**;
 - **SSH**.

В общем случае рекомендуется использовать протокол SSH. С точки зрения начинающего пользователя различие заключается в том, что при использовании протокола HTTPS каждый раз для соединения с удалённым репозиторием потребуется ввод логина и пароля, тогда как в случае с SSH — нет. При этом для того, чтобы использовать SSH необходимо провести первоначальные настройки. На самом деле, протокол SSH является предпочтительным по ряду технических причин среди которых можно выделить более высокий уровень безопасности, а также эффективное сжатие данных. Для подробного ознакомления с преимуществами и недостатками различных протоколов рекомендуется ознакомиться со [следующим официальным материалом \[27\]](#).

2.3.6.2. Настройка соединения с удалённым репозиторием посредством протокола SSH

Для установления связи с удалённым репозиторием GitHub посредством протокола SSH необходимо осуществить ряд действий, а именно генерировать пару SSH-ключей, а затем добавить их в профиль аккаунта на портале GitHub.

2.3.6.2.1. Проверка наличия существующих SSH-ключей. Для начала необходимо проверить наличие существующих ключей. Для этого следует запустить Терминал и ввести команду:

```
$ ls -al ~/.ssh
```

— в случае наличия существующих ключей Терминал возвратит примерно следующее сообщение:

```
> total 20
drwx----- 2 user.name user.name 4096 aug 14 14:42 .
drwxr-xr-x 36 user.name user.name 4096 sept 1 09:14 ..
-rw----- 1 user.name user.name 464 aug 11 11:05
    id_ed25519
-rw-r--r-- 1 user.name user.name 107 aug 11 10:04
    id_ed25519.pub
-rw-r--r-- 1 user.name user.name 1326 aug 11 19:08
    known_hosts
```

— в этом случае можно пропустить второй этап, описанный в подсекции [2.3.6.2.2–67](#), и перейти к третьему этапу, описанному в подсекции [2.3.6.2.3 на с. 67–68](#). В случае отсутствия существующей пары необходимо осуществить её генерацию.

2.3.6.2.2. Генерация новой пары ключей. Для создания пары ключей на основе алгоритма RSA, необходимо запустить Терминал и выполнить команду:

```
$ ssh-keygen -t rsa -b 4096 -C "user.name@host.com"
```

— указав при этом тот адрес электронной почты, который указан в профиле на GitHub. Помимо адреса электронной почты аргументами команды являются: алгоритм генерации ключа и его длина в битах. Следует сказать, что алгоритм RSA не является единственным. О различиях между алгоритмами RSA [99], DSA [85], ECDSA [86] и Ed25519 [14] можно почитать в следующих статьях и комментариях к ним: [22, 74, 69]. В целом, можно сказать, что для целей обучения анализу данных, равно как и для большинства практических целей оценщиков нет существенной разницы в том, какой алгоритм будет использован при создании пары ключей. Однако, с точки зрения соответствия лучшим практикам и современным тенденциям можно сказать следующее:

- a) Алгоритм DSA несколько устарел и подвержен уязвимости, поскольку решение проблемы вычислительной сложности взятия логарифмов в конечных полях [108], на которой он и был основан, было найдено в 2013 году.
- b) Схожий с DSA алгоритм ECDSA лишён указанного недостатка, поскольку основан не на конечном числовом поле [111], а на группе точек эллиптической кривой [128]. При этом криптографическая стойкость алгоритма в существенной степени зависит от возможности компьютера генерировать случайные числа.
- c) Алгоритм RSA обеспечивает достаточную надёжность при условии достаточной длины ключа. Длина ключа в 3072 либо 4096 бит является достаточной. Данный алгоритм является рекомендуемым в том случае, если нет возможности использовать алгоритм Ed25519.
- d) Алгоритм Ed25519 является предпочтительным во всех случаях, когда система технически способна работать с ним. Данный алгоритм обеспечивает хорошую криптостойкость, при этом работа с ключами происходит существенно быстрее, чем при использовании алгоритма RSA. Длина публичного ключа составляет всего 68 символов, тогда как RSA генерирует публичный ключ длиной в 544 символа (при 3072 бит).

Таким образом, вместо вышеуказанной команды рекомендуется использовать команду:

```
$ ssh-keygen -t ed25519 -C "user.name@host.com"
```

— адрес электронной почты также должен совпадать с тем, который указан в профиле на портале GitHub. Терминал возвратит сообщение:

```
> Enter a file in which to save the key (/home/user.name/.ssh/id_ed25519): [Press enter]
```

— предложив нажать Enter для сохранения ключей в каталоге по умолчанию. Следует согласиться с предложением и перейти к этапу создания пароля:

```
2307 > Enter passphrase (empty for no passphrase): [Type a
2308 passphrase]
2309 > Enter same passphrase again: [Type passphrase again]
```

2312 — ключи SSH готовы. Для возможности работы с ними необходимо добавить их в ssh-agent. Для этого сначала необходимо запустить ssh-agent в фоновом режиме, выполнив последовательно две команды:

```
2315 $ sudo -s -H
2316 $ eval "$(ssh-agent -s)"
```

2319 — далее осуществляется добавление самого ключа:

```
2320 $ ssh-add ~/.ssh/id_ed25519
```

2323 — ключи зарегистрированы в ssh-agent и могут быть использованы для взаимодействия с порталом GitHub.

2325 **2.3.6.2.3. Добавление публичного ключа на портал GitHub.** Для того, чтобы добавить в профиль на портале GitHub публичный ssh ключ необходимо получить его значение. Для начала следует установить xclip:

```
2328 $ sudo apt-get update
2329 $ sudo apt-get install xclip
```

2332 — теперь существует возможность автоматически копировать возвращаемые терминалом сообщения в буфер обмена. Сделаем это:

```
2334 $ xclip -selection clipboard < ~/.ssh/id_ed25519.pub
```

2337 — в данный момент буфер обмена содержит значение публичного ssh ключа. После этого необходимо зайти в свой профиль на портале GitHub и найти пункт меню **Settings**, а затем **SSH and GPG keys**. После этого следует нажать на кнопку **New SSH key**. Откроется меню, состоящее из двух полей: заголовка и значение ключа. В поле заголовка можно ввести любые символы, например имя, фамилию и должность. В поле значения ключа необходимо вставить содержимое буфера обмена. Существует семь возможных начальных символов ключа:

- 2344 ● «ssh-rsa»;
- 2345 ● «ecdsa-sha2-nistp256»;
- 2346 ● «ecdsa-sha2-nistp384»;
- 2347 ● «ecdsa-sha2-nistp521»;
- 2348 ● «ssh-ed25519»;
- 2349 ● «sk-ecdsa-sha2-nistp256@openssh.com»;

- «sk-«ssh-ed25519@openssh.com»

— в зависимости от применённого алгоритма. В случае совпадения практического значения с одним из возможных, можно сделать вывод о том, что все подготовительные операции были выполнены корректно. Нет необходимости вглядываться в имеющееся на практике значение: система в любом случае не зарегистрирует ключ, не отвечающий требованиям по маске. В том случае, если ключ прошёл валидацию, кнопка Add SHH key, расположенная ниже поля, станет активной. После её нажатия произойдёт добавление ключа.

Перед началом использования связи по SSH протоколу рекомендуется провести проверку. Для этого в терминале следует ввести команду:

```
$ ssh -T git@github.com
```

— терминал запросит ввести пароль, установленный при генерации ключей. В случае установления успешной связи терминал возвратит сообщение:

```
> Hi Kirill-Murashev! You've successfully authenticated, but
  GitHub does not provide shell access.
```

— связь установлена, возможна работа с удалённым репозиторием.

2.3.6.3. Создание и установка GPG ключа.

2.3.6.3.1. Основные сведения. Использование GPG ключей необходимо для подтверждения подлинности авторства commit. Использование подписи ключом GPG не является обязательным условием при работе с GitHub. Более того, в повседневной рутинной практике оценки чаще всего не возникает необходимость создания публичного репозитория и верификации commit. Однако с учётом возрастающих рисков киберугроз, усиления важности вопросов информационной безопасности, а также порой возникающей необходимости юридического доказывания авторства отчёта об оценке и подлинности его содержания, краткое изучение вопросов использования цифровой подписи представляется целесообразным. Весьма интересной выглядит история проекта. Его первоначальное название G10 является символической отсылкой к 10-й статье Конституции Федеративной Республики Германии [7], гарантирующей тайну переписки и связи. Наиболее известной программой, осуществляющей шифрование и подпись сообщений и файлов, стала PGP (Pretty Good Privacy) [95], разработанная в 1991 году Филиппом Циммерманом [68, 126]. В 1997 году был выпущен открытый стандарт OpenPGP. Его open-source реализацией стал GNU Privacy Guard (GnuPG или GPG) [31], разработанный в 1999 году Вернером Кохом [112].

Для начала, как и в случае с ключами SSH, ключи GPG (приватный и публичный) необходимо генерировать. GitHub поддерживает несколько алгоритмов генерации ключей:

- RSA

- 2392 • ElGamal
- 2393 • DSA
- 2394 • ECDH
- 2395 • ECDSA
- 2396 • EdDSA

2397 — рекомендуемым по умолчанию является алгоритм RSA&RSA 4096.

2398 **2.3.6.3.2. Проверка наличия существующих ключей.** Необходимо запустить
 2399 Терминал и использовать команду:

```
2400 $ gpg --list-secret-keys --keyid-format=long
2401
2402
```

2403 — либо, в зависимости от системы:

```
2404 $ gpg2 --list-keys --keyid-format LONG
2405
2406
```

2407 — во втором случае может потребоваться предварительная настройка, выполняемая
 2408 путём выполнения команды:

```
2409 $ gpg2 --list-keys --keyid-format LONG
2410
2411
```

2412 — в случае отсутствия пары ключей, следует перейти к шагу, описанному в под-
 2413 секции [2.3.6.3.3–70](#), в случае наличия данный шаг можно пропустить и перейти
 2414 к описанному в подсекции [2.3.6.3.4 на следующей странице–71](#).

2415 **2.3.6.3.3. Генерация пары ключей GPG** В Терминале следует ввести команду:

```
2416 $ gpg --gen-key
2417
2418
```

2419 — Терминал возвратит сообщение, предложив выбрать алгоритм:

```
2420 > Please select what kind of key you want:
2421 (1) RSA and RSA (default)
2422 (2) DSA and Elgamal
2423 (3) DSA (sign only)
2424 (4) RSA (sign only)
2425
2426
```

2427 — следует выбрать 1 либо 2. Далее терминал предложит выбрать длину ключа.
 2428 Рекомендуются использовать длину в 4096 бит в случае выбора пункта RSA&RSA
 2429 и 2048 в случае DSA&Elgamal. Далее следует указать срок действия пары ключей
 2430 либо поставить «0» для генерации бессрочных ключей. Данный выбор не является
 2431 необратимым: срок действия пары ключей возможно изменить впоследствии. Далее
 2432 необходимо указать данные пользователя и придумать пароль.

2433 После генерации пары следует проверить её существование путём использования
 2434 команды:

```
2435 $ gpg --list-secret-keys --keyid-format=long
2436
2437
```

2438 — либо, в зависимости от системы:

```
2439 $ gpg2 --list-keys --keyid-format LONG
2440
2441
```

2442 — Терминал возвратит примерно следующее сообщение:

```
2443 pub      dsa2048/169D4D0EC86C0000 2021-08-14 [SC]
2444
2445     ...
2446 uid                               [ultimate] kirill.murashev (my-key) <
2447     kirill.murashev@gmail.com>
2448
2449     ...
```

2450 — в данном случае идентификатором публичного ключа является значение «169D4D0EC86C0000».
2451 Введём команду:

```
2452 $ gpg --armor --export 169D4D0EC86C0000
2453
2454
```

2455 — Терминал возвратит полное значение публичного ключа, начинающееся с:

```
2456 -----BEGIN PGP PUBLIC KEY BLOCK-----
2457
2458
```

2459 — и заканчивающееся:

```
2460 -----BEGIN PGP PUBLIC KEY BLOCK-----
2461
2462
```

2463 — полученное значение необходимо скопировать, после чего можно перейти к следу-
2464 ющему шагу. Дополнительные сведения о работе с GPG можно получить по [ссыл-](#)
2465 [ке \[30\]](#).

2466 **2.3.6.3.4. Добавление публичного ключа на портал GitHub.** Необходимо зай-
2467 ти на портал GitHub. В меню **Settings** выбрать пункт **SSH and GPG keys**, далее
2468 нажать **New GPG key**, вставить значение публичного ключа из буфера обмена
2469 и нажать **Add GPG key**. При выполнении последнего действия система предло-
2470 жит ввести пароль от аккаунта.

2471 Теперь существует возможность создавать подписанные commit. Для подписи
2472 конкретного commit следует использовать дополнительные аргумент **-S** команды
2473 **git commit**. Пример такой команды:

```
2474 $ git commit -S -m "commit_message"
2475
2476
```

2477 — при этом система потребует ввести пароль, придуманный при генерации пары
2478 ключей. Для включения глобальной опции подписания всех commit по умолчанию
2479 следует ввести команду:

```
2480 $ git config --global --edit
2481
2482
```

2483 — в открывшемся окне текстового редактора установить следующие значения:

```
[user]
    name = <User.Name>
    email = <user.name@host.com>
    signingkey = <169D4D0EC86C0000> [gpg]
    program = gpg2 [commit]
    gpgsign = true
```

— значения, заключённые в <>, естественно должны быть своими.

2.3.6.4. Установление связи между локальным и удалённым репозиториями

2.3.6.4.1. Отправка содержимого локального репозитория в удалённый.

В первую очередь необходимо скопировать ссылку на репозиторий из меню **Code**. После этого следует зайти в каталог локального репозитория, запустить из него Терминал и ввести команду:

```
$ git remote add origin <hyperref>
```

— указав вместо «hyperref» конкретную ранее скопированную ссылку на удалённый репозиторий.

Для просмотра настроек удалённого репозитория следует ввести команду:

```
git remote -v
```

— терминал возвратит, например такое сообщение:

```
origin https://github.com/Kirill-Murashev/
      AI_for_valuers_book.git (fetch)
origin https://github.com/Kirill-Murashev/
      AI_for_valuers_book.git (push)
```

— как видим имеет место существование двух репозиториях:

- **fetch** служит для чтения содержимого удалённого репозитория;
- **push** — для отправки содержимого локального репозитория в удалённый.

Для отправки данных на удалённый сервер следует применить команду:

```
git push origin master
```

— где:

- **push** — указание на действие, которое необходимо выполнить;
- **origin** — наименование сервера, на который следует отправить данные;
- **master** — название ветки, в которую необходимо отправить данные.

Содержимое локального репозитория в том состоянии, в котором оно было зафиксировано в последнем commit, отправлено в удалённый репозиторий на портале GitHub. При этом создаётся дополнительный указатель `remotes/origin/master`, называемый веткой слежения. Данный указатель следует для хранения данных о том, на каком commit находится указатель `Head` в ветке `master` на удалённом сервере `origin`.

2.3.6.4.2. Получение содержимого удалённого репозитория. Для получение содержимого удалённого репозитория на локальный компьютер необходимо выбрать каталог, в который планируется загрузка и запустить из него Терминал. При этом следует иметь ввиду, что в данном каталоге будет сформирована новая папка, имя которой будет повторять имя удалённого репозитория — источника. Далее следует использовать команду `git clone`, аргументом которой будет являться ссылка на удалённый репозиторий. Следующие команды предназначены для создания локальной копии репозитория полезных, а зачастую и необходимых для изучения данного материала:

```
$ git clone git@github.com:Kirill-Murashev/  
AI_for_valuers_book.git
```

— создание локальной копии исходного кода данного руководства, его версии в формате PDF, а также дополнительных материалов, использованных при создании.

```
$ git clone git@github.com:Kirill-Murashev/  
AI_for_valuers_R_source.git
```

— создание локальной копии репозитория, содержащего код на языке R, предназначенного для выполнения процедур, описанных в данном руководстве.

```
$ git@github.com:Kirill-Murashev/  
AI_for_valuers_Python_source.git
```

— создание локальной копии репозитория, содержащего код на языке Python, предназначенного для выполнения процедур, описанных в данном руководстве.

2.3.6.4.3. Обновление репозитория. В процессе работы особенно в случае совместной работы нескольких специалистов над одним проектом возникает необходимость частой синхронизации локальных репозитория разработчиков. Для выполнения обновления содержимого локального репозитория следует зайти в его локальный каталог, запустить Терминал и использовать команду `git pull`, в качестве аргументов которой указываются имена сервера и ветки:

```
git pull origin master
```

— Git загрузит изменения в случае их наличия.

Поскольку данный проект активно развивается, автор рекомендует выполнять обновление репозитория, содержащих текст данного руководства и программный код, не реже одного раза в месяц.

В процессе совместной работы нескольких специалистов может возникнуть ситуация, при которой они оба захотят отправить свои изменения на сервер. В Git предусмотрена защита: разработчик, отправивший свои изменения позже, получит сообщение об ошибке и предложение выполнить `git pull` в том случае, если его изменения конфликтуют с изменениями первого разработчика, т. е. сервер не может выполнить процедуру `fast forward`. Во избежание такой ситуации рекомендуется всегда сначала использовать команду `git pull`, обновляющую данные о том, на каком `commit` находится указатель `Head` на сервере, и загружающую изменения. Для обновления данных и перемещения ветки слежения без загрузки новых `commit` с сервера можно использовать команду:

```
git fetch origin
```

— а затем использовать команду:

```
git merge origin/master
```

— использование последовательности этих команд равнозначно использованию одной команды `git pull`.

2.3.7. Работа с Git в IDE

2.3.7.1. Работа в RStudio

End

2.3.7.2. Работа в Spyder

End

2.3.7.3. Работа в PyCharm

End

End

2.3.8. Заключение

Данный раздел содержал лишь основные сведения и инструкции по работе с Git и Github, достаточные для первичной настройки и начала работы. Для более подробного ознакомления с Git и Github можно порекомендовать просмотр данного [видеоурока \[133\]](#), а также изучение [официального руководства \[4\]](#).

2.4. Установка и настройка

2.4.1. Git

2.4.1.1. Установка на операционных системах, основанных на Debian: Debian, Ubuntu, Mint и т. п.

В операционных системах, основанных на ядре Linux [66], относящихся к ветке Debian [93], Git зачастую бывает уже установлен вместе с системой. Чтобы проверить наличие Git в командную строку терминала следует ввести:

```
git
```

В случае наличия Git в системе, терминал возвратит длинное сообщение, начинающееся примерно следующим образом:

```
usage: git [--version] [--help] [-C <path>] [-c <name>=<
value>]
[--exec-path[=<path>]] [--html-path] [--man-path] [--info-
path] [-p | --paginate | -P | --no-pager] [--
no-replace-objects] [--bare] [--git-dir=<path
>] [--work-tree=<path>] [--namespace=<name>]
```

В случае его отсутствия:

```
Command 'git' not found, did you mean:
```

Во втором случае следует использовать следующие команды:

```
sudo apt update -y
sudo apt install git -y
```

Процесс проходит автоматически и не требует внимания со стороны пользователя.

2.4.1.2. Установка на операционной системе Windows

Установка Git на Windows осуществляется обычным для данной операционной системы образом. Необходимо загрузить установочный файл с соответствующей страницы [24] и запустить процесс установки, желательно приняв при этом все настройки по умолчанию.

2.4.1.3. Установка на macOS

Существует несколько способов установки Git на macOS. Их перечень приведён на соответствующей странице [25] сайта Git. Следует отметить, что в случае наличия в системе Xcode [129] Git также уже присутствует, и его установка не требуется. В данном материала приводится один из возможных способов. Для начала необходимо установить менеджер пакетов Homebrew [34]. Для этого в командной строке терминала необходимо ввести следующую команду:

```
/bin/bash -c "$(curl -fsSL https://raw.githubusercontent.com
/Homebrew/install/HEAD/install.sh)"
```

После этого можно перейти к установке самого Git. Для этого в командной строке терминала необходимо ввести следующую команду:

```
brew install git
```

Как и в случае, описанном выше в секции [2.4.1.1 на предыдущей странице–74](#), процесс проходит автоматически и не требует внимания со стороны пользователя.

2.4.2. R

2.4.2.1. Установка на операционных системах, основанных на Debian: Debian, Ubuntu, Mint и т. п.

2.4.2.1.1. Установка на операционных системах Ubuntu, Mint и производных от них. Как правило для установки достаточно зайти в Центр приложений, ввести в строку поиска «CRAN» и установить R посредством графического интерфейса. Однако, в зависимости от дистрибутива есть вероятность получения относительно устаревшей версии. Для получения сведений о текущей версии R следует зайти на [официальный сайт \[81\]](#) и узнать там номер и дату последнего релиза. На момент написания данных строк таковой является версия 4.1.1 (Kick Things) от 2021-08-10. Для проверки версии, установленной в системе в Терминале следует ввести команду:

— в случае автора терминал возвратил сообщение:

```
R version 4.1.1 (2021-08-10) -- "Kick Things"
Copyright (C) 2021 The R Foundation for Statistical
Computing
Platform: x86_64-pc-linux-gnu (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and 'citation()'
on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.
```


— версия соответствует последнему релизу. В большинстве случаев установка и использование не самой последней версии не вызывает никаких проблем. Однако, в случае, если есть стремление использовать самых свежий стабильный релиз следует отказаться от установки через Центр приложений и выполнить следующую последовательность команд. Сначала необходимо добавить доверенный ключ:

```
$ sudo apt-key adv --keyserver keyserver.ubuntu.com --recv-keys
E298A3A825C0D65DFD57CBB651716619E084DAB9
```

— система возвратит следующее сообщение:

```
Executing: /tmp/apt-key-gpghome.cul0ddtmN1/gpg.1.sh --
keyserver keyserver.ubuntu.com --recv-keys
E298A3A825C0D65DFD57CBB651716619E084DAB9
gpg: key 51716619E084DAB9: public key "Michael Rutter <
marutter@gmail.com>" imported
gpg: Total number processed:
1 gpg: imported: 1
```

— ключ добавлен, можно добавить репозиторий:

```
$ sudo add-apt-repository 'deb https://cloud.r-project.org/
bin/linux/ubuntu focal-cran40/'
```

— далее обновляем зависимости и устанавливаем R:

```
$ sudo apt update -y
$ sudo apt install r-base
```

— R установлен и готов к использованию.

2.4.2.1.2. Установка на операционной системе Debian. Одной из основных особенности операционной системы [Debian](#) [10] является её стабильность и надёжность. В особенности это касается её ветки stable. Однако достоинства и недостатки часто являются продолжением друг друга. Многие приложения, доступные из стандартных репозиториях Debian могут быть представлены в версиях, отстающих от актуальных на 0.5–2 года. Таким образом, в случае использования операционной системы Debian ветки stable рекомендуется провести самостоятельную установку актуальной версии R. Следует выполнить последовательность команд:

```
$ sudo apt install dirmngr --install-recommends
```

— либо, в случае её недоступности:

```
$ sudo apt install software-properties-common
```

— обе эти команды добавляют необходимый в дальнейшем инструмент `add-apt-repository`. Далее устанавливаем инструмент, необходимый для обеспечения работы протокола `https` при передаче данных из репозитория:


```
2738 $ sudo apt install apt-transport-https
2739
2740
```

2741 Далее добавляем доверенный ключ:

```
2742 $ sudo apt-key adv --keyserver keys.gnupg.net --recv-key '
2743 E19F5F87128899B192B1A2C2AD5F960A256A04AF '
2744
2745
```

2746 — система возвратит сообщение:

```
2747 Executing: /tmp/apt-key-gpghome.y6W4E0Gtftp/gpg.1.sh --
2748 keyserver keys.gnupg.net --recv-key
2749 E19F5F87128899B192B1A2C2AD5F960A256A04AF
2750
2751 gpg: key AD5F960A256A04AF: 4 signatures not checked due to
2752 missing keys gpg: key AD5F960A256A04AF: public key "
2753 Johannes_Ranke_(Wissenschaftlicher_Berater)<johannes.
2754 ranke@jrwb.de>" imported
2755 gpg: Total number processed: 1
2756 gpg: imported: 1
2757
```

2758 — теперь можно перейти к установке самого R. Следует обратить внимание на тот факт,
2759 что содержание аргумента приведённой ниже команды зависит от используемой
2760 версии OS Debian. На момент написания этих строк текущей стабильной версией
2761 является Debian 11 «bullseye». В этом случае команда будет выглядеть следующим
2762 образом:

```
2763 $ deb http://cloud.r-project.org/bin/linux/debian bullseye -
2764 cran40/
2765
2766
```

2767 — для получения дополнительных сведений следует обращаться к [соответствующей](#)
2768 [странице](#) [11] сайта R. Далее следует выполнить последовательность команд:

```
2769 $ sudo apt update -y
2770 $ sudo apt install r-base
2771
2772
```

2773 — R установлен и готов к использованию.

2774 2.4.2.2. Установка на операционных системах Windows и macOS

2775 В данном случае установка не требует никаких специфических действий и осу-
2776 ществляется путём загрузки установочного файла с [соответствующей страницы](#)
2777 сайта R [81] и запуска установщика.

2778 2.4.3. RStudio

2779 Независимо от используемой операционной системы самым простым способом
2780 установки RStudio является загрузка установочного образа, соответствующего опе-
2781 рационной системе, со [страницы](#) сайта RStudio [77].

2782 2.4.4. Python

2783 End

2784 2.4.5. Spyder

2785 End

2786 2.4.6. PyCharm

2787 End

2788 2.4.7. SQL

2789 End

2790 End

2791 Теоретическая часть

2792 Глава 3.

2793 Математическая основа анализа
2794 данных

2795 End

Глава 4.

Основные понятия

4.1. Что было раньше: курица или яйцо? Соотношение понятий **statistics, machine learning, data mining, artificial intelligence**

На сегодняшний день можно говорить о существовании множества понятий, описывающих применение математических и статистических методов при решении практических задач. В целом, можно без преувеличения сказать, что в настоящее время нет такой области деятельности человека, в которой бы не применялись математические методы и модели. Невозможно охватить все аспекты применения математических методов — в данном разделе будут рассмотрены лишь интересующие нас вопросы анализа данных применительно к оценке. Существует несколько общепринятых понятий, описывающих группы методов и подходов, применяемых при анализе данных и укоренившихся в сознании общества. В таблице 4.1.1 приводится перечень наиболее распространённых терминов.

Таблица 4.1.1. Перечень понятий, описывающих группы методов анализа данных

№	Англоязычный термин	Русскоязычный термин
0	1	2
1	Statistics	Математическая статистика
2	Machine Learning	Машинное обучение
3	Data mining	Интеллектуальный анализ данных
4	Artificial intelligence	Искусственный интеллект

Из таблицы 4.1.1 следует, существуют как минимум, четыре разных понятия, описывающих широкую, но всё же единую с точки зрения конечной цели область. Целью данного раздела является попытка разобраться в следующих вопросах:

- что представляет из себя каждое направление;
- что есть общего у них, и в чём они различаются;

- какие именно методы и средства используются в каждом из этих направлений;
- чем мы будем заниматься в процессе изучения данной работы.

Следует отметить, что на сегодняшний день в вопросе того, как именно следует разделять эти понятия, отсутствует консенсус. На эту тему продолжают вестись дискуссии. Забегая вперёд, можно сказать, что скорее всего нет смысла говорить о жёстком разделении этих понятий. Едва большинство конкретных методов могут применяться в рамках каждого из этих направлений. И всё же, по мнению автора, вопрос соотношения вышеуказанных понятий заслуживает должного внимания. Следует отметить, что большая часть рассуждений и выводов, сделанных в данном разделе, является не более чем мнением автора и не должна рассматриваться как-то иначе.

Вряд ли требуется много слов для того, чтобы объяснить, что представляет собой **математическая статистика**. В целом, можно сказать, что данный раздел математики тесно связан с Теорией вероятности и использует единый с ней понятийный аппарат. При этом, математическая статистика допускает как **частотный** [127], так и **байесовский** [105] подход к понятию вероятности. В целом, можно сказать, что методы математической статистики, основанные на частотной вероятности, основываются на свойствах данных (например распределениях), исходят из базовой предпосылки о случайности распределения значений переменных, случайности каких-либо различий между выборками и отсутствии значимых зависимостей между переменными и предназначены для построения интерпретируемых моделей, описывающих взаимосвязь между данными. Как правило, применение методов статистики начинается со спецификации модели, методы которой, в общем случае, позволяют вывести конкретную модель на основе минимизации функции потерь, неизбежных вследствие наличия «шума» [75]. Сильной стороной математической статистики является хорошая интерпретируемость результатов, а также возможность применения в условиях ограниченного числа наблюдений. Слабой стороной является зависимость от распределения значений переменных, возможность работы только с *данными, характеризующими отношения*. Классическим примером применения методов математической статистики является построение линейной регрессионной модели.

Машинное обучение представляет построение алгоритмов, способных обучаться. Обучением алгоритма (программы) на основе опыта считается такой процесс, при котором по мере обучения производительность этого алгоритма (программы) возрастает в соответствии с некоей производительности. Основным направлением машинного обучения является обучение по прецедентам. Автоматический сбор и распознавание исходных данных также относятся к задачам машинного обучения. При этом многие методы машинного обучения напрямую взяты из математической статистики [75]. Машинное обучение находится на стыке математической статистики, методов оптимизации и классических математических дисциплин, но также имеет собственную специфику, связанную с проблемами **вычислительной эффективности** [55] и **переобучения** [54] [51]. Примером применения методов машинного обучения является построение случайного леса решающих деревьев.

Интеллектуальный анализ данных в узком смысле, предложенном в 1992 году Г. И. Пятецким-Шапиро [118], представляет собой процесс обнаружения в сырых данных ранее неизвестных, нетривиальных, практически полезных, доступных интерпретации знаний, необходимых для принятия решений в различных сферах человеческой деятельности [50]. В широком смысле интеллектуальный анализ данных представляет собой применение человеком методов математической статистики и машинного обучения применительно к конкретным данным для решения конкретных задач. Как правило, речь идёт о большом и сложном наборе данных. Целью применения методов интеллектуального анализа данных чаще всего является получение знаний в тех областях, в которых их не хватает либо предсказание точных значений будущих наблюдений. При этом, как правило, нет цели формирования выводов о свойствах данных либо эта цель второстепенна [75]. Концепция интеллектуального анализа данных базируется на следующих предпосылках:

- данные могут быть неточными, неполными (содержать пропуски), противоречивыми, разнородными, косвенными, и при этом иметь гигантские объёмы, вследствие чего их интерпретация в конкретном случае требует значительных интеллектуальных усилий;
- сами алгоритмы анализа данных могут обладать элементами машинного «интеллекта», в частности, способностью обучаться по прецедентам, т. е. общие выводы на основе частных наблюдений, при этом разработка подобных алгоритмов также требует значительных интеллектуальных усилий;
- процессы переработки сырых данных в структурированные, структурированной в информацию, а информации в знания уже не могут быть выполнены вручную, и требуют нетривиальной автоматизации [50].

Понятие **искусственный интеллект** представляет собой область деятельности, включающую в себя науку и технологии, направленную на создание искусственных интеллектуальных агентов, т. е. алгоритмов и компьютерных программ, способных действовать так, как мог бы действовать настоящий интеллектуальный агент, например человек. Искусственный интеллект не обязательно должен основываться на индуктивных методах. Например, он может быть просто настроен на увеличение значений некой переменной X всеми доступными способами безотносительно конкретных применяемых при этом алгоритмов. В определённом смысле любые успешные действия машины, осуществляемые ей одновременно не детерминировано и не случайным образом, можно назвать искусственным интеллектом [75, 56].

Следует отметить, что данное понятие очень популярно в массовой культуре, искусстве и общественном сознании, вследствие чего подвержено неверному толкованию. Современное понятие искусственного интеллекта, а также современное состояние науки о нём не предполагают создание искусственного интеллекта равного человеческому либо хоть сколько-нибудь сопоставимого с ним. Более того,

в случае необходимости проведения параллелей с биологическими интеллектуальными агентами, правильнее говорить не об искусственном интеллекте, а об искусственных инстинктах и рефлексам. Стремление минимизировать функцию ошибки при построении модели линейной регрессии можно сравнить с инстинктом, а попытку перебора её коэффициентов — с рефлексом.

В определённом смысле можно сказать, что правильнее говорить не о создании интеллектуальных агентов (вопрос того, что такое интеллект, чем он отличается от сознания, разума, воли — является слишком сложным и выходит за рамки науки о искусственном интеллекте), а о создании рациональных агентов, т. е. таких, которые обеспечивают принятие рациональных, а по возможности оптимальных, решений на основе фактически доступных данных [75].

Подытоживая всё вышесказанное можно повторить, что чёткое разграничение этих четырёх понятий нецелесообразно, а скорее всего невозможно. Ниже приводятся краткие выводы, сделанные автором в ходе изучения вопроса соотношения данных понятий:

Математическая статистика — наука и совокупность практических методов, нацеленные на количественный анализ данных, описание их свойств и построение моделей. Основывается на вероятностных моделях. При этом использование компьютера не является обязательным.

Машинное обучение — область деятельности, нацеленная на создание алгоритмов способных обучаться, а также оптимизацию вычислений. При этом конечной целью данных алгоритмов является предсказание значений новых наблюдений. Как правило, не фокусируется на свойствах данных. Чаще основывается на байесовском подходе к вероятности. Использование компьютера является обязательным, при этом важную роль играет вычислительная эффективность.

Интеллектуальный анализ данных — область деятельности, направленная на получение новых знаний и объяснение закономерностей в конкретной области путём применения методов статистики и машинного обучения. Фокусируется на практическом применении, оптимизации наборов данных. Использование компьютера является обязательным.

Искусственный интеллект — область деятельности, направленная на создание рациональных агентов, способных применять любые доступные методы для формирования рациональных решений на основе всех доступных данных. Компьютер сам является агентом.

В таблице 4.1.2 на с. 86 приведены обобщающие сведения. Возникает вопрос: чем именно мы будем заниматься в процессе работы с данным руководством. С одной стороны, как уже было сказано выше, нельзя провести жёсткое разграничение между всеми этими понятиями. С другой, каждое из этих направлений деятельности имеет свою специфику. Кроме того, нельзя забывать о том, что целью данной работы

является применение современных технологий анализа данных в конкретной области деятельности — оценке стоимости. Для начала можно сделать краткий обзор уже существующих в российской практике разработок в каждом направлении.

Математическая статистика в российской оценочной практике применяется сравнительно давно. Существует некоторый набор литературы и методических материалов, посвящённых применению её методов в оценке. Можно вспомнить ряд работ С. В. Грибовского, Н. П. Барина, В. Г. Мисовца, а также иных авторов, посвящённых вопросам применения математических методов в оценочной деятельности. Основным предметом интереса данных работ чаще всего выступает применение корреляционного и регрессионного анализа в их параметрических вариантах. Как правило в них рассматривается применение программного продукта Microsoft Excel в качестве основного и единственного инструмента анализа. Не погружаясь в рассуждения о методах, предлагаемых в данных работах, их достоинствах и недостатках, в целом, можно высказать некоторое сожаление о том, что данное направление развития оценки в принципе не получает широкого распространения, почти полностью уступая место практикам оценки, основанным на применении минимального числа наблюдений, называемых в терминологии оценки объектами-аналогами, и последующей корректировки значений на основе неких «справочников». Таким образом, можно сказать, что применение методов математической статистики в оценке не является чем-то принципиально новым. Данная работа также будет включать существенный объём материала, посвящённого их применению. Как уже было сказано выше, одним из достоинств математической статистики является хорошая интерпретируемость результатов, что является важным в условиях необходимости их доказывания и защиты. Ещё одной сильной стороной является развитый аппарат описательных методов, позволяющий делать общие выводы о свойствах данных, в нашем случае — свойствах открытых рынков. Таким образом, рассмотрение и применение методов математической статистики представляется необходимым.

Машинное обучение пока что не получило широкого распространения в среде профессиональных оценщиков. При этом уже существует ряд сервисов, позволяющих предсказывать стоимость объектов на основе методов машинного обучения. Примером является [Калькулятор недвижимости \[131\]](#), созданный Циан.Групп. По мнению автора, необходимость применения методов машинного обучения обусловлена в частности тем, что существенная часть данных открытых рынков не относится не только к *данным, характеризующим отношения*, но в принципе не представляет собой количественные данные. Адрес, конструктивные особенности, техническое состояние — всё это качественные данные, с трудом поддающиеся какой-либо трансформации в количественные. Существуют методы квалиметрии, позволяющие провести преобразование качественных данных в количественные, однако в любом случае на выходе могут быть получены лишь порядковые данные, применение к которым методов математической статистики возможно лишь в ограниченном объёме. При этом многие методы машинного обучения свободны от зависимости от типа исходных данных, а также вида их распределения. Кроме того, существует необходимость автоматизации сбора исходных данных. Ситуация, при которой оценщики либо их помощники вручную копируют данные с сайтов объявлений, а затем вруч-

ную вставляют в свои рабочие материалы, выгляди нелепой для 2021 года.

Интеллектуальный анализ данных представляет собой междисциплинарное направление деятельности. Очевидно, что планирование исследования, выбор источников данных, поиск и объяснение закономерностей, соотнесение промежуточных и итоговых результатов с априорными знаниями, их интерпретация и описание, — всё это является важной частью деятельности оценщика.

Искусственный интеллект не имеет однозначного и точного определения. Любой недетерминированный алгоритм, позволяющий решать практические задачи неслучайным образом может считаться реализацией искусственного интеллекта.

Таким образом, все четыре направления представляют интерес и будут рассматриваться в данной работе. Можно лишь в очередной раз повторить мысль о том, что любое их разделение носит приблизительный и отчасти условный характер.

Таблица 4.1.2. Обобщение сведений

	Математическая статистика	Машинное обучение	Интеллектуальный анализ данных	Искусственный интеллект
Приоритетный подход к вероятности	Частотный	Байесовский	Оба	Байесовский
Опирается на свойства данных	Да	Нет	Да	Нет
Требовательность к количеству наблюдений	Умеренная	Высокая	Высокая	Средняя
Применение компьютера	Желательно	Обязательно	Обязательно	Обязательно
Основной инструмент	R	Python	R и Python	Python
Включает в себя методы подготовки эксперимента и отбора данных	Нет	Нет	Да	Нет
Предполагает наличие знаний в области, в которой проводится эксперимент	Нет	Нет	Да	Нет
Стремится к вычислительной эффективности	Нет	Да	Да	Да
Позволяет делать общие выводы о свойствах данных	Да	Нет	Да	Нет
Позволяет давать точные прогнозы значений новых наблюдений	Частично	Да	Да	Да
Основная область знаний, к которой относится направление	Математика	Информатика	Междисциплинарная	Информатика
Обеспечивает хорошую интерпретацию результатов	В большинстве случаев	86/88 Не всегда	10 сентября 2021 г. Да	Нет

²⁹⁹⁵ Глава 5.

²⁹⁹⁶ Начало работы с R

Рабочая версия

2997 Глава 6.

2998 Автоматизированный сбор данных

2999 The End