DOI: 10.14515/monitoring.2015.2.04
УДК 303.432:004.6

В. Смирнов

НОВЫЕ КОМПЕТЕНЦИИ СОЦИОЛОГА В ЭПОХУ БОЛЬШИХ ДАННЫХ

НОВЫЕ КОМПЕТЕНЦИИ СОЦИОЛОГА В ЭПОХУ «БОЛЬШИХ ДАННЫХ»

СМИРНОВ Владимир Алексеевич - доктор социологических наук, доцент, проректор ФГБОУ ВПО Костромской ГСХА. E-mail: kano_igt@mail.ru

Аннотация. Описана современная социокультурная ситуация эскалации «больших» и открытых данных, проанализированы ситуации, в которых социологу могут потребоваться новые компетенции, такие как программирование и кодинг. Говоря о ситуации возникновения большого объема данных, автор указывает на возможность возникновения двух ключевых исследовательских стратегий: самолокализации и погружения в данных. Первая предполагает концентрацию учёного на традиционных формах социологического анализа, ограниченного небольшими выборочными исследованиями, вторая же ориентирована на максимальное использование все возрастающего объема открытых данных.

При работе с большими и открытыми данными неизбежно возникают следующие задачи, требующие решения: 1. извлечение данных; 2. Их обработка и придание им «опрятного» вида; 3. визуализация данных. Поэтому исследователю необходимо овладение такими компетенциями как кодинг и программирование. Автор полагает, что для социологов с гуманитарным образованием наиболее подходят статистическая среда R и язык программирования Python, они имеют достаточно понятный синтаксис и позволяют решать сложные задачи по обработке, анализу и визуализации данных.

В статье приведены примеры реально

SOCIOLOGIST`S NEW COMPETENCES IN THE TIMES OF “BIG DATA”

SMIRNOV Vladimir Alekseevich - Doctor of Sociology, Senior Lecturer, Vice-Rector, Kostroma State Agricultural Academy. E-mail: kano_igt@mail.ru

Abstract. The article describes the modern escalation of “big” and open data; situations where sociologist needs new competences such as coding and programming are analyzed. The author puts attention to two key research strategies when dealing with big data: self-localization and dipping into data. The first strategy refers to the researcher’s focus on traditional sociological analysis restricted by small sample size; the second one implies maximal use of the growing open source data.

The challenges a researcher has to face when dealing with big data are as follows: (1) data extraction; (2) data processing and improvements; (3) data visualization. This is why a researcher should master coding and programming. Sociologists with humanitarian education should choose R statistical environment and Python programming language becasue they have simple syntax and allow solving complicated data processing, analysis and visualization tasks.

The article provides examples of scripts as a demonstration of the possibilities of the R and Python programming languages.

работающих скриптов, с целью демонстрации возможностей языков программирования R и Python для извлечения, анализа и визуализации данных.

Ключевые слова: BigData, большие данные, открытые данные, кодинг, программирование, компетенции, статистическая среда R, язык программирования Python

Keywords: BigData, open data, coding, programming, competences, R statistical computing environment, Python programming language

Отечественная социология находится в ситуации своеобразной методологической травмы, которая проявляется в «растерянности исследователей перед обилием социологических теорий, методологий, методов в процессе принятия решений о выборе средств познавательной деятельности» [Татарова, 2006]. Такая травма проявляется и в отсутствии четкой рефлексии относительно собственной методологической стратегии, «аргументированных различий при трактовке базовых теоретико-методологических оснований, затрагивающих принципиальные вопросы, прежде всего предмета и объекта социологии» [Тощенко, 2014].

Сегодня наблюдаются очертания травмы, обусловленной эскалацией информации и данных. Все чаще встает проблема больших данных (Big Data), требующая осмысления и интерпретации. Big Data не имеет четких и признаваемых всеми дефиниций, тем не менее можно выделить ее ключевые особенности: это не только «возросший объем, но и возросшая скорость передачи данных и разнообразие источников» [Фрэнкс, 2014]. Перечислим главные отличия больших данных от традиционных.

Во-первых, большие данные часто автоматически генерируются машиной без участия человека, тогда как традиционные всегда предполагают его присутствие. Возьмем, к примеру, розничные или банковские транзакции, запись телефонных звонков, доставку товаров или выставление счетов на оплату. Все эти действия подразумевают присутствие человека. Кто-то должен внести деньги, сделать покупку или платеж, позвонить по телефону, отправить посылку. В каждом случае частью процесса создания новых данных остается человек. С большими данными дело обстоит иначе. Например, встроенный в двигатель датчик генерирует их в автоматическом режиме.

Во-вторых, они нередко соотносятся с новыми источниками данных. Например, использование почасовых данных с фитнес-трекеров пользователей позволяет по-новому взглянуть на практики здорового образа жизни, ориентируясь не на мнения респондентов и не на количество продаж, а на транслирование в режиме реального времени различных показателей — от количества пройденных шагов до часов, выделяемых на сон.

В-третьих, многие источники больших данных не создавались как дружественные к пользователю. Возьмем, к примеру, текстовые потоки от сайта социальных медиа. Пользователей невозможно убедить соблюдать все правила грамматики, синтаксиса или лексические нормы. Когда люди публикуют запись, исследователь получает колоссальный объем текстовых данных, которые невозможно проанализировать вручную. Это ставит вопрос о необходимости овладения навыками кодинга и создания скриптов для анализа, в том числе неструктурированного текста [Фрэнкс, 2014].

Сегодня социологи все чаще используют большие данные для анализа социальных институтов, практик, норм и т.д. Вот лишь некоторые примеры: исследование городских

трущоб на основе данных сотовых операторов [Eagle, 2010], анализ сообщений Twitter и прогнозирование на их основе социальных настроений [Golder, 2011], анализ читательских предпочтений на основе чтения данных с электронных устройств для чтения книг [Alter, 2012].

Мы будем использовать в связке с термины «большие» и «открытые» данные, под ними мы понимаем машиночитаемые наборы данных, которые выкладываются в сеть государственными, муниципальными, общественными структурами, сообществами гражданских интернет-активистов. Так, на портале открытых данных России (http://data.gov.ru/) на момент написания статьи было собрано 2419 наборов данных в машиночитаемых форматах.

Проблемность ситуации заключается не только в возрастающем объеме новых данных, но и в их псевдодоступности. Исследователь ежедневно может видеть в Интернете тысячи релевантных ссылок, открытые машиночитаемые данные, данные, интегрированные в код страниц сайтов. Все они доступны, но требуют такой кропотливой ручной работы по извлечению и анализу.

Попытаемся выделить ключевые направления совершенствования компетенций социолога. Основным, по-нашему мнению, является умение эффективно работать с данными с момента их извлечения до визуализации. Для этого исследователю необходимо овладеть навыками программирования и работы в статистических средах, с интерфейсом командной строки. Постараемся показать, как всего несколько строчек программного кода могут решать сложные, с точки зрения обработки больших массивов данных, задачи.

Профессиональные стратегии социолога

В современной ситуации у исследователя есть две ключевые стратегии профессионального выживания. **1. Самолокализация.** Предполагает жесткое самоограничение в получаемых, обрабатываемых, используемых данных. Многие социологи, исповедующие философию полевика-качественника, выбирают именно ее, уходя в миры case study. При этом методологическая ориентация исследователя не является ключевым фактором при выборе этой стратегии, многие ученые, верящие в количественную социологию, также ориентированы на нее. Ограничиваясь узким сегментом собранных лично ими или их коллегами данных, используя результаты социологических исследований крупных научных и исследовательских центров, они воспроизводят социологические практики самолокализаторов. Характерная черта группы — нежелание или неумение работать с «сырыми», «неопрятными» большими данными, заполонившими интернет-ресурсы и требующими владения инструментами их «огранки». **2. Погружение в данные.** Предполагает наличие не только широкого спектра собственно профессиональных компетенций, но и запас компетенций надпрофессиональных, таких, как кодинг, программирование, Data Mining. Реализация этой стратегии требует от социолога программного перевооружения. Исследователям приходится осуществлять своеобразный инструментальный возврат (не в техническом плане, а скорее в идеологическом, аксиологическом), в основе которого лежит философия работы с данными «своими руками». Возникает потребность отказаться от «кнопочной» работы, заложенной в таких статистических пакетах, как SPSS, Statistica т.д.

Не всегда эффективна и ситуация аутсорсинга, когда деятельность по сбору и обработке данных не осуществляется социологами, ученые лишь интерпретируют данные, адаптированные под решаемые ими задачи. Каковы причины этого? Помимо традиционных

аргументов, ориентирующих социолога на профессиональную рефлексию не только в сфере проектирования выборки, особенностей социологического инструментария, но и на четкое понимание, как работает тот или иной статистический метод, есть еще одно обстоятельство. Дело в том, что мощные «кнопочные» инструменты анализа данных, которые в большинстве случаев используются сегодня, сами по себе ригидны: каждый новый релиз, например, SPSS, содержит какие-либо обновления как в интерфейсе программы, так и в методах анализа данных, но скорость обновления таких программ не может сравниться со скоростью появления новых пакетов статистической среды R, предлагающих исследователю современные методы и алгоритмы анализа. В условиях, когда новые математические и статистические алгоритмы появляются с такой же скоростью, с какой растет объем открытых данных, это обстоятельство играет важную роль. Значимо и то, что все «кнопочные» программы платные, что в ряде случаев не позволяет исследователю иметь на персональном компьютере лицензированные релизы тех или иных программных статистических комплексов.

Аналитические инструменты социолога, погруженного в данные

Как уже отмечалось, работа с большими и открытыми данными ведет к необходимости технического перевооружения исследователя. Одно из его направлений — переход со статистических пакетов с «кнопочным» интерфейсом на аналитические инструменты, основанные на интерфейсе командной строки, предполагающей наличие у ученого знаний кодинга и программирования. В качестве привлекательного и популярного аналитического инструмента можно назвать статистическую среду и язык программирования R. Эта аналитическая платформа распространяется бесплатно и имеет много приверженцев, постоянно совершенствующих как ее программную оболочку, так и количество и объем доступных методов анализа данных. С каждым годом растет количество книг и статей, посвященных R, на английском и русском языках, что позволяет исследователю самостоятельно осваивать новые пакеты и методы анализа (см., например, http://r-analytics.blogspot.ru/2013/10/r.html#.VL-bdiyqDlo). Существует также большое количество Интернет-ресурсов, посвященных этому проекту [http://r-analytics.blogspot.ru].

Основные преимущества R:

- большое количество инструментов по адаптации и преобразованию данных под запрос исследователя;
- возможность видеть результаты манипуляций с данными, которые осуществляют те или иные статистические алгоритмы;
- возможность настраивать среду для решения конкретных специфических задач в той или иной области исследований;
- большое число модулей и пакетов, позволяющих использовать самые современные методы анализа данных;
- возможность самостоятельно создавать скрипты, модули, пакеты, позволяющие решать ту или иную задачу в сфере обработки и анализа данных;
- возможность напрямую считывать машиночитаемые данные с Интернет-ресурсов и работать с ними, как с обычными данными;
- большая библиотека графических методов, позволяющих упаковывать данные в привлекательные и легко интерпретируемые рисунки и диаграммы;

– интеграция с текстовыми и издательскими системами, такими, как LaTex, Microsoft Word, позволяющая создавать аналитические отчеты по итогам исследования непосредственно в среде R, используя ее язык программирования.

Как показывает практика работы с открытыми данными (и в машиночитаемых форматах, и в просто встроенных в веб-страницы), исследователь может столкнуться со спектром проблем по управлению данными, которые невозможно решить средствами R. Возникает потребность не просто в освоении статистических платформ, имеющих встроенный язык программирования, а в использовании специализированного языка, благодаря которому можно извлекать и трансформировать данные. На наш взгляд, наиболее приемлем для решения социологических задач язык Python. Это язык высокого уровня, позволяющий писать простые, но функциональные и эффективные скрипты по адаптации и преобразованию данных. С помощью Python легко работать с онлайн-ресурсами. Ключевое преимущество данного языка программирования — понятный и легко воспринимаемый синтаксис, простота освоения.

Язык программирования Python в связке со средой R — уникальный набор инструментов: схожесть этих языков, особенно при функциональном программировании, известная простота освоения их базовых, необходимых для исследователя структур, дают социологу возможность быть эффективным исследователем в эпоху больших данных.

Ключевые задачи

Полагаем, что социолог, работающий в парадигме «погружение в данные» должен владеть навыками решения следующих задач.

I. Извлечение данных из разных виртуальных источников, их скачивание в уже пригодном для анализа формате. Несмотря на широкое распространение данных в машиночитаемых форматах, может потребоваться самостоятельно извлечь данные с того или иного сайта и превратить их в машиночитаемый формат. Эту задачу можно решить несколькими строками скрипта. Специализированная библиотека языка — Beautiful Soup позволяет получать код html и вытаскивать из него релевантные данные. Важно понимать, что необходимые для анализа данные могут находиться на нескольких десятках, а то и сотнях страниц, для просмотра которых может потребоваться несколько дней кропотливой работы. Овладение навыками написания небольших скриптов, извлекающих информацию из того или иного Интернет-ресурса, становится эффективным инструментом повышения продуктивности работы.

В качестве примера приведем скрипт, написанный в рамках нашего исследовательского проекта, в котором проанализированы особенности институционализации малых инновационных предприятий (МИП), созданных российскими университетами [Смирнов, 2015]. На сегодняшний день более 400 вузов в стране создали от одного до нескольких десятков МИПов, каждый из которых имеет в своем активе разное число интеллектуальных продуктов. Если объединить все эти признаки (вуз, название МИПа, интеллектуальный продукт, вид интеллектуального продукта, регион создания МИПа) в матрицу «объект–признак», получится таблица, имеющая более 3000 записей. Собирать их вручную трудно; нескольких строк кода на языке Python позволяет решить эту задачу за несколько

минут. Ниже приводится фрагмент кода, написанного нами парсера (программа для сбора данных с Интернет-ресурса) для сайта, на котором находится реестр МИПов (https://mip.extech.ru/reestr.php), чтобы продемонстрировать простоту решения указанной задачи.

```
def spis (url):
 #Парсер извлекает название вуза и мипа
 response=urllib.request.urlopen (url)
 soup=BeautifulSoup (response)
 nam1 = soup.find_all ('td', class_='dark')[1].text.strip ()
 pro1=[]
 for i in soup.find_all ('tr', class_='fild'):
 cols=i.find_all('td')
 pro1.append ({
 'mip': cols [0].text.strip(),
 'vuz': nam1,
 })
 for i in soup.find_all ('tr', class_='line'):
 cols=i.find_all('td')
 pro1.append ({
 'mip': cols [0].text.strip(),
 'vuz': nam1,
 })
 return pro1
```

Приведенный фрагмент кода позволяет извлечь с сайта названия вуза и МИПа. В теле парсера есть отдельные функции (строки кода), позволяющие извлекать остальные признаки объектов, перемещаться по страницам сайта и записать полученные результаты в файл с машиночитаемым форматом (csv). Представленный код базируется на простых конструкциях, которые могут быть освоены социологом в короткий срок. Это, во-первых, циклы, имеющиеся в любом языке программирования (в нашем случае цикл «for»), использование изменяемых последовательностей, таких как списки, и наконец, функции поиска «find_all». Создание парсера под каждый конкретный сайт — стандартная задача, в рамках которой меняется содержание кода, но общий алгоритм остается неизменным. Представим ключевые шаги этого процесса.

1 Нахождение в коде сайта раздела, представляющего необходимые для исследователя данные. Решение этой задачи, например, при использовании браузера Mozilla Firefox, возможно с помощью дополнения Firebug.
2 Определение ключевых тегов в коде, репрезентирующих необходимые исследователю данные.
3 Загрузка этих тегов вместе с данными с использованием библиотеки языка Python –Beautiful Soup.
4 Поиск в загруженном контенте релевантных данных, их запись в список и файл в машиночитаемом формате.

Овладение навыком извлечения данных из Интернет-ресурсов позволяет социологу быть мобильным, эффективнее решать интересующие задачи, быстрее проверять гипотезы.

II. Обработка данных и придание им «опрятного» вида. Концепция «опрятных» данных (Tidy Data) активно разрабатывается сторонниками статистической среды R [Hadley, 2014], для чего в ней создано несколько «пакетов» (например, tidyr), позволяющих оптимизировать их. Два наиболее важных свойства таких данных состоят в следующем:

- каждый столбец в таблице соответствует одной переменной;
- каждая строка соответствует одному наблюдению.

Используя «опрятные» данные и соответствующие программные инструменты, исследователь тратит больше времени на изучение свойств данных, а не на их обработку. Несмотря на наличие специализированных пакетов в среде R, иногда возникает потребность решить ряд специфических задач. Здесь мы снова можем использовать преимущества и простоту Python. Одна из задач — «расширение числа признаков объекта», «вытаскивание» из данных дополнительной информации. Например, имеется машиночитаемый список членов избирательных комиссий России всех уровней, включающий в себя Ф.И.О., должность, наименование политической или общественной организации, направившей члена комиссии. Общее число записей — более 760 000. Для обработки такого массива данных вручную потребуется работа десятков человек и несколько дней. Скрипт на Python позволяет решить эту задачу за 5 минут. Приведем пример скрипта, позволяющего определить пол члена избиркома и добавить его в качестве отдельного столбца в матрицу «объект–признак».

```
at=open («fi1.txt»)
zx=open («fi_pol.txt», «w»)
for i in (at):
 if i[-3]==«a»:
 i=i.replace («\n», «,»)
 zx.write (i+ «женский\n»)
 else:
 i=i.replace («\n», «,»)
 zx.write (i+ «мужской\n»)
at.close()
zx.close ()
```

Несколько простых строк кода существенно облегчают жизнь исследователю, повышая качество первоначальных данных. Представленный скрипт делает несколько важных вещей. Он открывает файл, в котором находится список членов избирательных комиссий, запускает цикл, при каждой итерации которого проверяет букву, на которую заканчивается отчество человека, а затем добавляет еще один признак — пол (если отчество заканчивается на букву «а», значит, речь идет о женщине). Такой подход позволяет обработать многотысячный список за пару минут, добавив к имеющимся данным в машиночитаемом формате новый признак,

который можно использовать для анализа особенностей гендерного состава избирательных комиссий России.

Приведем еще один пример. Сегодня социологи все чаще используют возможности онлайн-опросов и для этого прибегают к созданию специализированных форм в среде Google. Такой подход удобен тем, что позволяет, распространив ссылку на анкету, получить ответы уже в машиночитаемом формате (csv). Пожалуй, единственная проблема здесь - неудачная обработка вопросов, имеющих несколько вариантов ответов. При получении данных в машиночитаемом формате эти ответы записываются в один столбец, и социолог должен привести их в «опрятный» вид, например, в вид дихотомических ответов, где выбранный вариант кодируется как «1», а невыбранный как «0». Такое представление множественных ответов позволяет социологу быть более вариативным при анализе вопроса, используя факторный анализ, методы регрессионного анализа и т.д. Для больших выборок ручное перекодирование таких вопросов требует немалых затрат. Мы попытались решить эту проблему с помощью небольшой функции, написанной на языке программирования Python, перекодировав множественные ответы в дихотомический формат. Поскольку данный скрипт является универсальным, приведем его полностью.

```
def perekod (num, file1, file2):
 ««««Функция превращает переменные в дихотомические»»««
 #file1 — это файл, где находятся данные
 #file2 — это файл куда записываются перекодированные данные
 fil1=open (file1)
 fil2=open (file2,»w»)
 print («Если переменные будут браться из файла, нажмите-1\n»
 «Если переменные будут вводиться вручную, нажмите-2»)
 a=int(input ())
 if a==1:
 file3=input («Введите имя файла (без кавычек): «)
 fil3=open (file3)
 z=0
 c=[«name»]*num
 for i in fil3:
 i=i.rstrip («\n»)
 i=i.rstrip (« «)
 i=i.rstrip («,»)
 c[z]=i
 z=z+1
 fil3.close()
 elif a==2:
 z=0
 c=[«name»]*num
 while z < num:
 c[z]=input («Введите значение переменной (без кавычек): «)
 z=z+1
 for i in fil1:
```

```
x=0
b=[«name»]*num
while x < num:
if c[x] in i:
b[x]=(«1» + «,»)
x=x+1
else:
b[x]=(«0» + «,»)
x=x+1
fil2.writelines(b)
fil2.write («\n»)
fil1.close ()
fil2.close ()
```

III. Визуализация данных. Время примитивных графических решений Excel прошло. При этом ни SPSS, ни SAS, ни другие статистические пакеты с визуальным интерфейсом не позволяют создавать многофункциональные, «заточенные» под конкретные данные диаграммы, графики, рисунки. Такие программные пакеты, как ggplot2, lattice дают возможность визуализировать даже сложные для наглядного представления данные. Важно, что в среде R имеются инструменты для создания нестандартных форматов графического представления данных, таких, как облако тегов, картограммы и т.д. Благодаря статистической среде R можно создавать инфографику.

В R имеются инструменты для создания не только привлекательных визуальных образов, но и диаграмм. В качестве примера приведем так называемые «диаграммы Кливленда». Они представляют собой графики, на которых точки-маркеры используются для отображения значений какой-либо количественной переменной (или переменных), разбитых на группы в соответствии с уровнями некоторой номинальной переменной (или переменных). Учёный показал, что столбиковые диаграммы (привычные для российских научных публикаций), используемые для изображения сгруппированных значений количественных переменных, плохо воспринимаются людьми [Cleveland, 1984]. Столбики были заменены точками, что позволило лучше видеть расстояния и различия между разными уровнями того или иного признака. Представим диаграмму Кливленда по итогам исследования, посвященного проблемам взаимодействия университетов и региональных сообществ. На ней обозначены затраты региональных органов власти на исследования и разработки по регионам ЦФО за 2013 г.

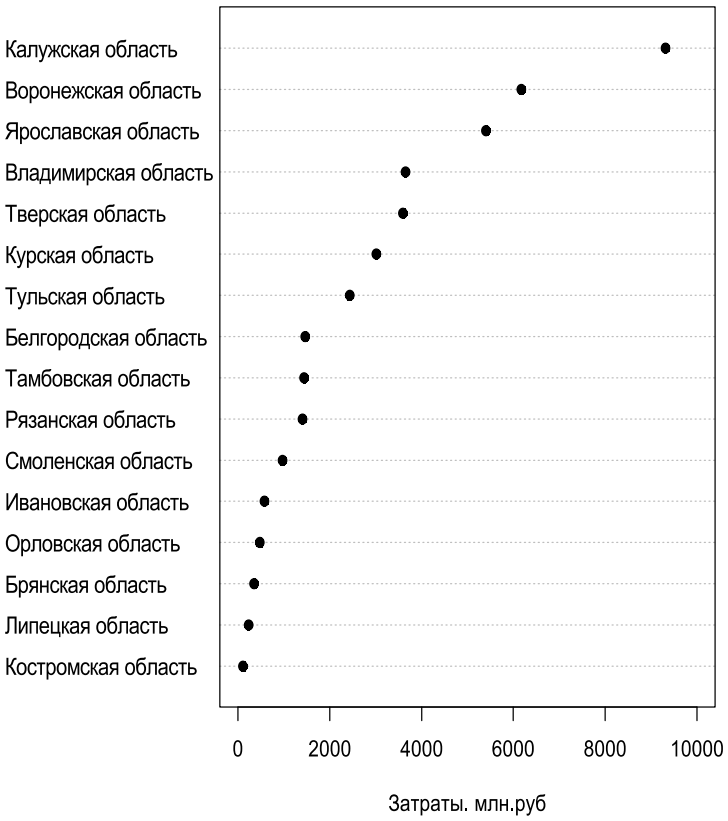

Рисунок 1 - Внутренние затраты на исследования и разработки по регионам ЦФО [Смирнов, 2014]

Представленная точечная диаграмма может быть создана в R одной строкой кода, который приведен ниже.

```
dotchart (X2013, labels=reg, pch=19, xlab=«Затраты. млн.руб», xlim=c(0,10000))    [1]
```

Код может легко изменяться, добавляя или модернизируя те или иные элементы диаграммы.

Для создания диаграммы требовалось загрузить файл в машиночитаемом формате в R, но факт того, что диаграммы можно создать в среде R, используя минимальный код, подтверждает эффективность и привлекательность для исследователей этого продукта. Отметим, что функция для создания точечных диаграмм имеет уже три версии, каждая из них включает дополнительные возможности по более эффективной визуализации данных.

Освоение социологом новых форм и практик работы с данными позволяет ему стать эффективным не только в интерпретации социальной реальности, но и в сборе информации о ней. Наряду с развитием профессиональной и методологической рефлексии, освоением новых теорий и конструктов, социолог может часть профессионального времени посвятить развитию надпрофессиональных компетенций, таких, как кодинг, программирование, работа с «большими данными».

Литература

1 Alter A. Your E-book Is Reading You // WSJ. 2012. June 29.

2 Cleveland W. S., Mc Gill R. Graphical perception: theory, experimentation, and application to the development of graphical methods// Journal of the American Statistical Association. 1984. 79(387): 531–554.
3 Eagle N. Big Data, Global Development and Complex Systems // Santa Fe Institute. 2010. 5 May.
4 Golder, Scott A. Diurnal and Seasonal Mood Vary with Work, Sleep and Day length Across Diverse Cultures // Science. Vol. 333, no. 6051. September 30. 2011. p. 1878–1881.
5 Hadley W. Tidy data //The Journal of Statistical Software. 2014. vol. 59.
6 Смирнов В.А. Интеграция университета в региональные процессы: возможные стратегии и ключевые факторы риска // Университетское управление: практика и анализ. 2014. № 6. С. 57–68.
7 Смирнов В.А. Малые инновационные предприятия российских вузов: особенности становления и ключевые противоречия // Социология образования. №3. 2015. С. 21–35.
8 Татарова ГГ. Методологическая травма социолога. К вопросу об интеграции знания // Социологические исследования. 2006. № 9. С. 1–18.
9 Тощенко Ж.Т. О социологических стратегиях в социологии // Вестник РГГУ, серия «Социологические науки». 2014. № 4. С. 11–21.
10 Фрэнкс Б. Укрощение больших данных. Как извлекать знания из массивов информации с помощью глубокой аналитики. М., 2014.