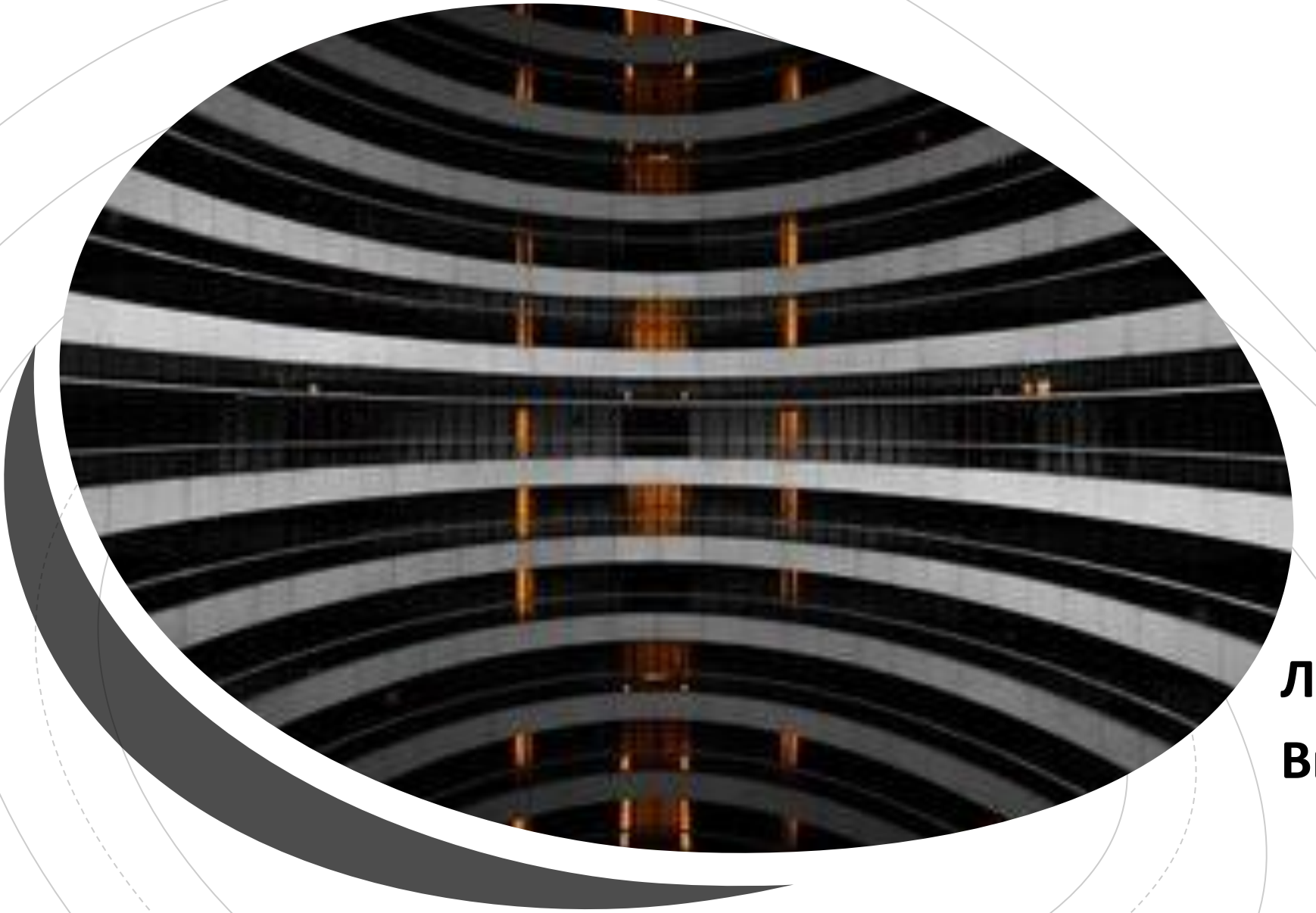


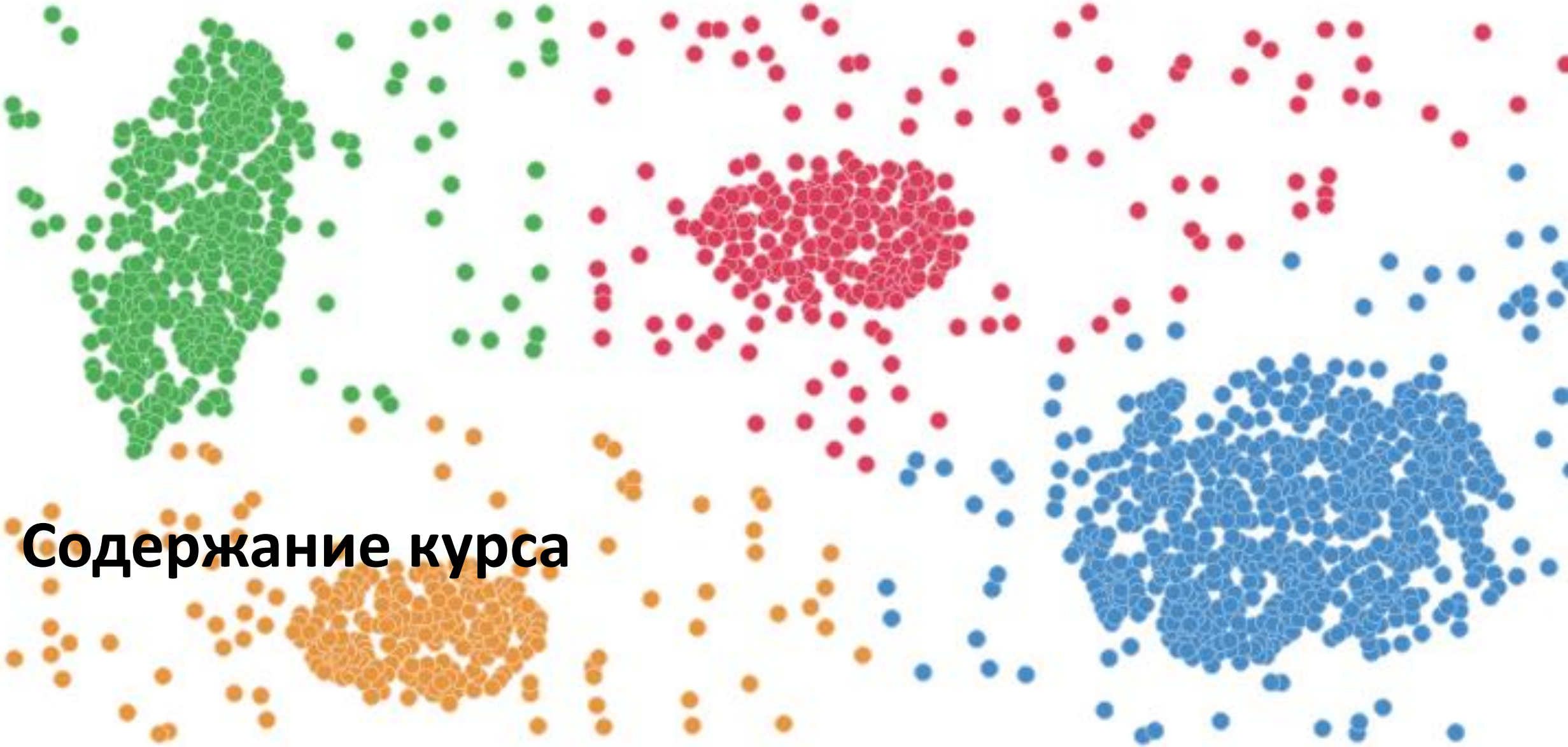
Smart data

Лекция #1

Введение.

И.А. Куликов

i.a.kulikov@gmail.com



Содержание курса

Содержание курса Smart data (Интеллектуальные данные)

- **Лекции:**

1. Введение, Дополнительные материалы.
2. Графы знаний.
3. Графы знаний (продолжение).
4. Практическое использование SPARQL.
5. *(на самостоятельную проработку) Big data и хранилища больших данных.*

- **Лабораторные работы.**

Smart Data. Определение

Smart Data. Определение.

Smart Data - это такая технология обработки данных, которая позволяет интегрировать различные источники данных (включая Big Data), устанавливать взаимосвязи между данными в этих источниках, анализировать данные из различных источников совместно. Цель Smart Data – обеспечить процессы принятия решений и операционную деятельность.

Fernando Iafrate, From Big Data to Smart Data, First published:27 February 2015, Print ISBN:9781848217553 |Online ISBN:9781119116189 |DOI:10.1002/9781119116189, Copyright © 2015 John Wiley & Sons, Inc.

Многие данные являются «большими» (с точки зрения объема, скорости и т. д.), но насколько они «Smart», то есть имеют значение для бизнеса?

Умные данные следует рассматривать как набор технологий и процессов, а также связанные с ним структуры (Центры компетенции бизнес-аналитики (BICC)), которые включают все значения, взятые из данных. Такие центры могут создаваться в организациях (корпорациях) и предназначены для определения задач, ролей, обязанностей и процессов для поддержки и продвижения эффективного использования бизнес-аналитики (BI) в организации.

Strange, K. H., Hostmann, B. (22 July 2003), BI Competency Center Is Core to BI Success, Gartner Research

Smart Data. Альтернативное определение.

Smart Data - это данные, в которые добавлено явное семантическое содержание посредством формализация метаданных (по определению, характеристике или процессу моделирования). Термин Smart в широком смысле говорит о том, что Smart (интеллект, сообразительность) - это измеримая величина по ее степени, другими словами, существуют степени яркости, точности, аккуратности, структуры и абстракции, в которой данные могут быть формально описаны.
Определение намеренно широко в смысле, что оно применимо как к сырым, обрабатываемым приложениями, к метаданным, к моделям (данные, бизнес-процессы и другие), и даже к метамоделям.
Smart Data - это продукт тщательного и открытого процесса, который описывает актуальную информацию, используемую предприятием. Полезная информация включает в себя данные, описывающие события, явления, материалы, процессы, процедуры, действия, приложения, структуры, отношения или сами данные.

James A. Rodger, Smart Data: Enterprise Performance Optimization Strategy (Wiley Series in Systems Engineering and Management), 2012

Smart Data. Важные замечания (1 из 3)

*Если соотносить определение **Smart Data** с известными **V**s из определения Big Data (объем, скорость, разнообразие и достоверность), то **Smart Data** связаны с **Достоверностью (Veracity)** наряду с еще одним часто используемым V:* ***Ценностью (Value).***

Используя интеллектуальные данные, мы ориентируемся на ценные данные и часто меньшие наборы данных, которые можно превратить в данные для обеспечения действий или операций предприятия, и в эффективные результаты для решения проблем клиентов и бизнеса. Речь идет об анализе и интерпретации данных, чтобы мы могли сделать процесс принятия решений и бизнес-функции предприятия основанными на данных, путем помещения их в контекст целей предприятия.

Smart Data. Важные замечания (2 из 3)

Smart Data *- это большие данные, которые превращаются в данные, которые можно использовать для действий, которые доступны в режиме реального времени для различных бизнес-результатов, будь то промышленные приложения, маркетинг на основе данных или оптимизация процессов.*

Используя интеллектуальные данные, мы действительно ищем способы избавиться от шума самого аспекта объема, так же как быстрые данные относятся к элементу скорости. Например, в контексте маркетинга и обслуживания клиентов интеллектуальные данные в основном рассматриваются с точки зрения гиперперсонализации.

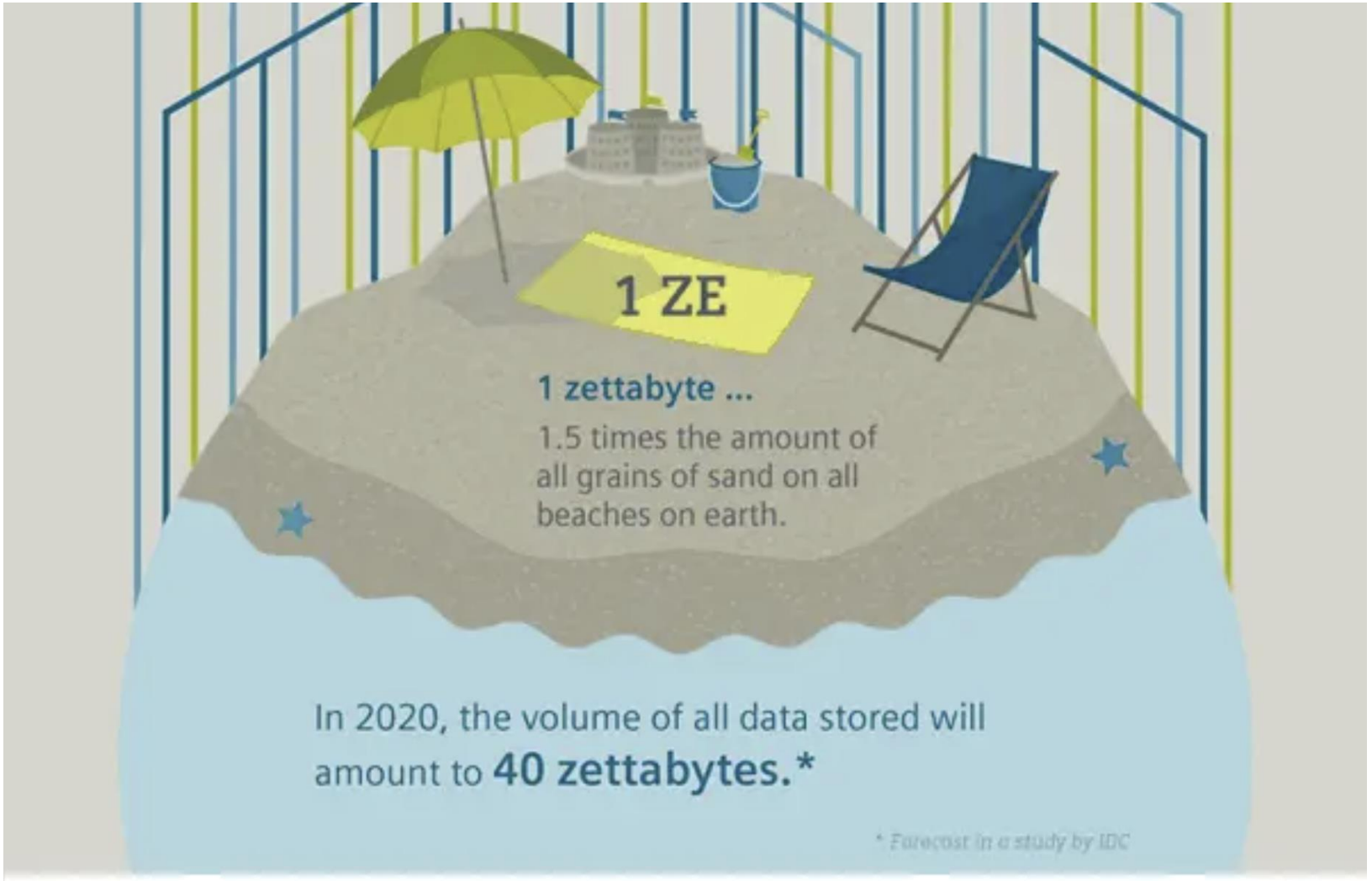
Smart Data. Важные замечания (3 из 3)

*Повышенное внимание к **Smart Data**, а не к **Big Data**, тесно связано с грядущей экономикой алгоритмов. Искусственный интеллект все чаще используется в бизнес-приложениях и для работы с **Big Data** и Интернета вещей **(IoT)**.*

Большинство имеющихся данных представляют собой неструктурированные данные, и только с помощью искусственного интеллекта и аналитики неструктурированные данные можно превратить в интеллектуальные данные и данные, пригодные для осуществления деятельности предприятия.

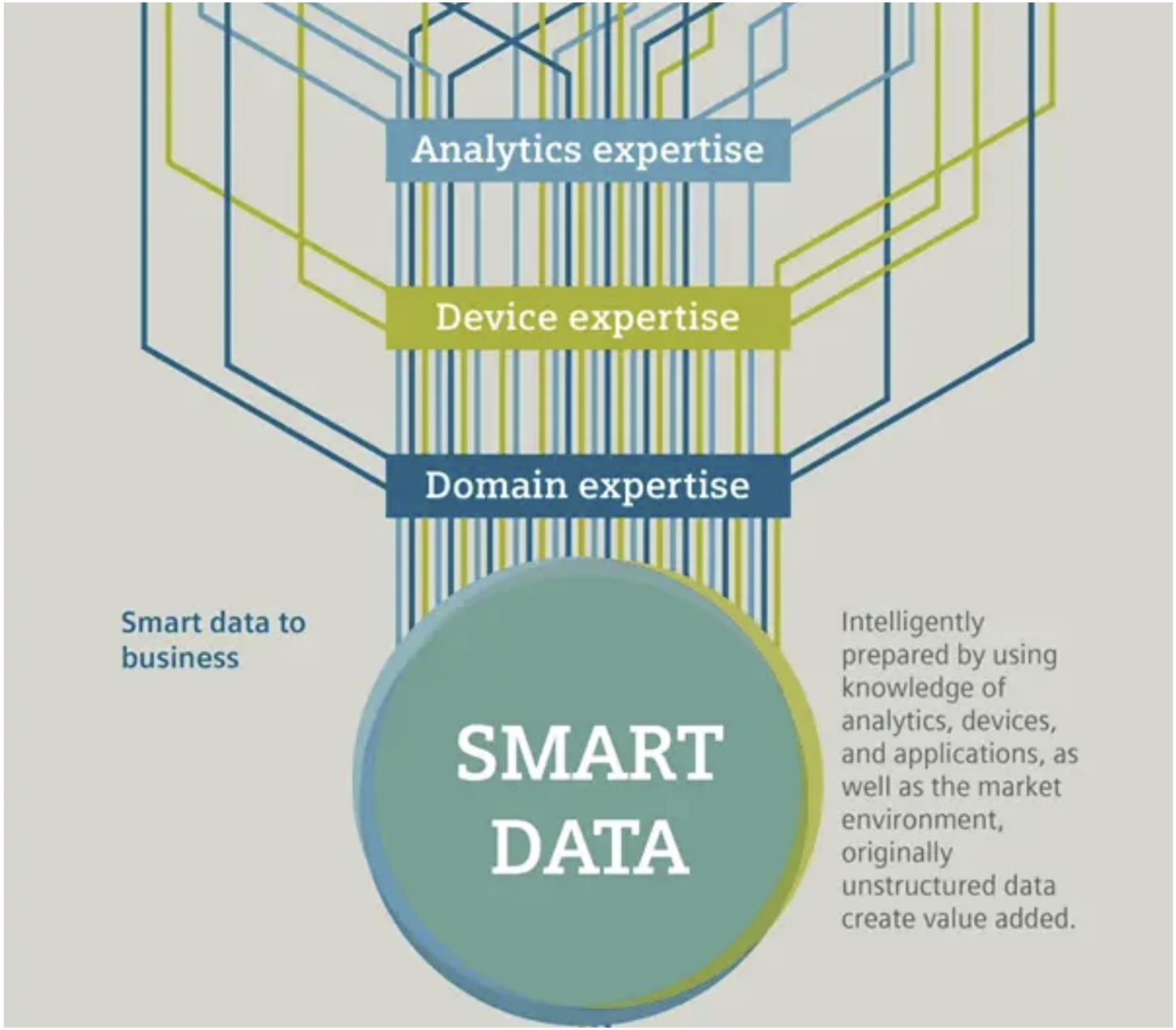
*Существенный вопрос в постоянно растущем объеме данных заключается в том, как использовать эти объемы на практике, и без аналитики, интерпретации и алгоритмов это невозможно. Приведенная ниже инфографика, разработанная **Siemens** для промышленных приложений, показывает проблему объема данных и необходимость перехода от больших данных к интеллектуальным данным.*

From Big Data to Smart Data (1 из 4)

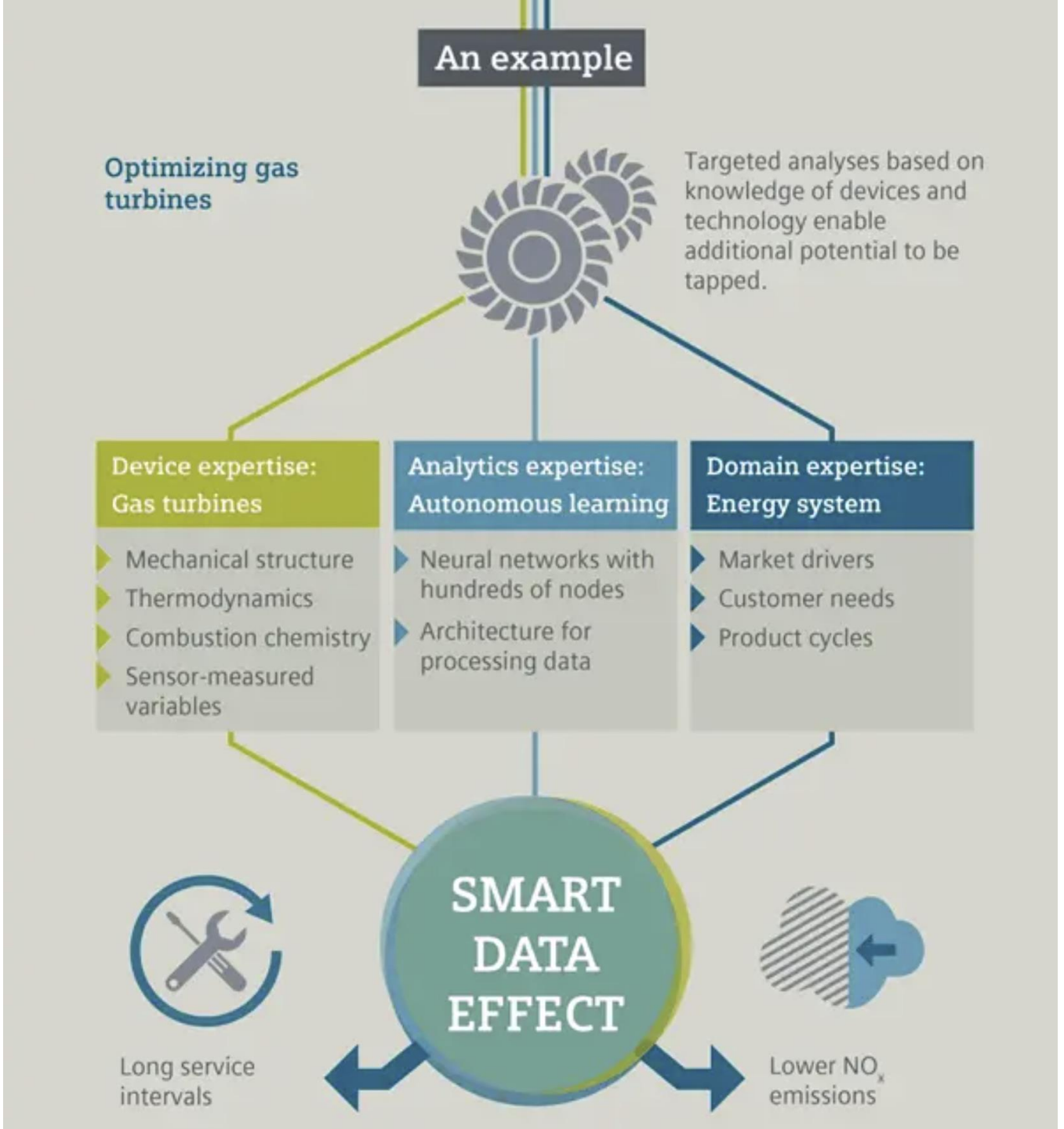


How can we make practical use of these data volumes?

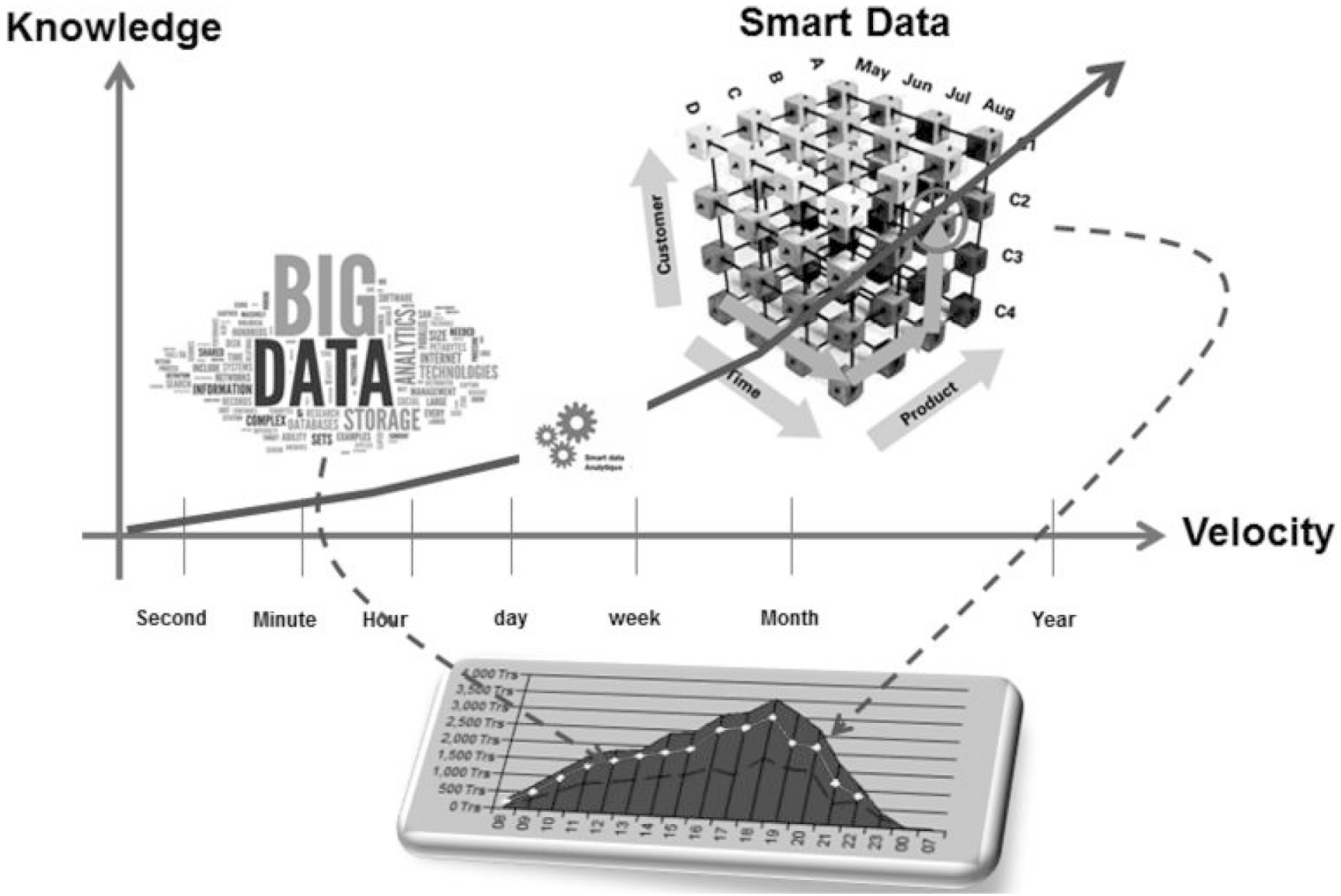
From Big Data to Smart Data (2 из 4)

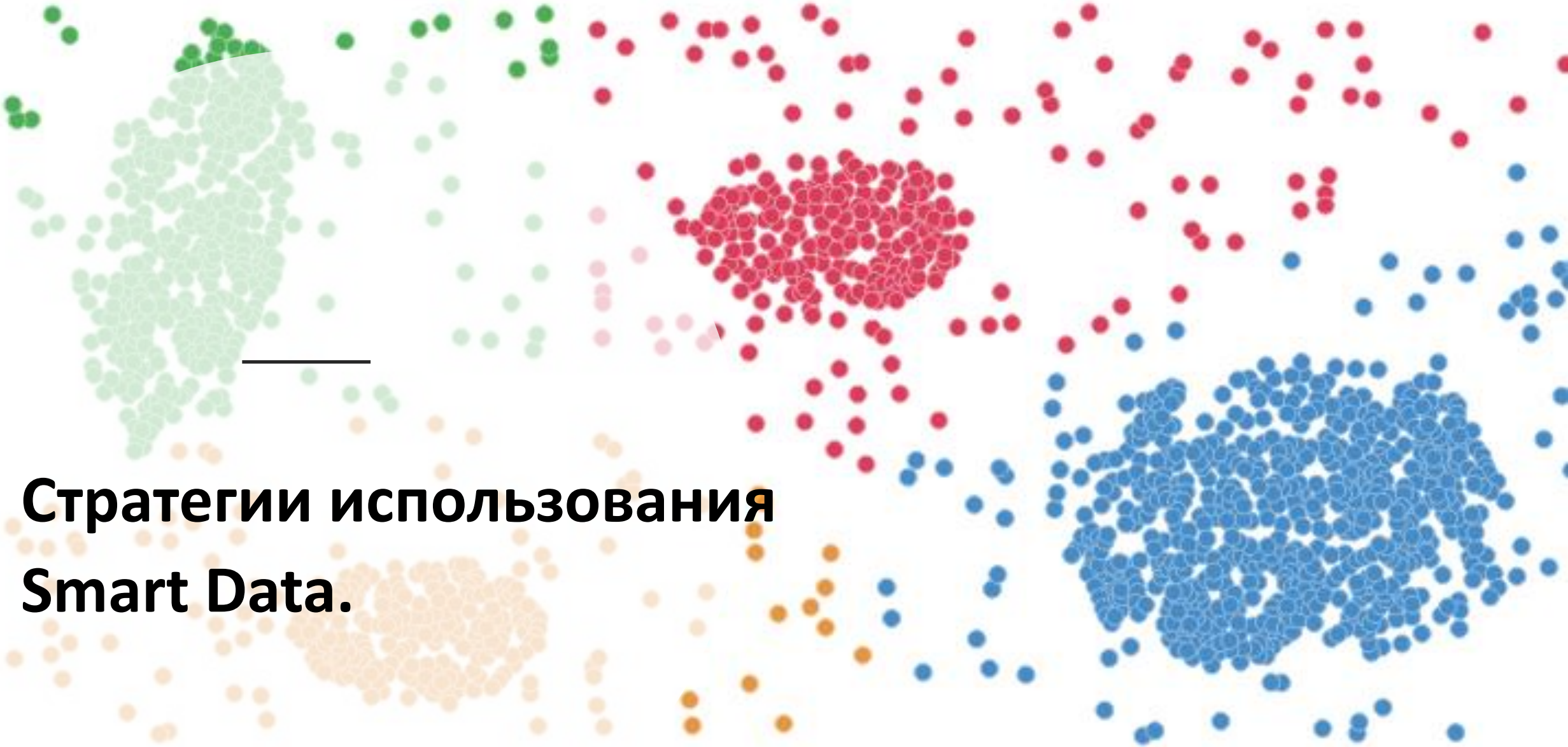


From Big Data to Smart Data (3 из 4)



From Big Data to Smart Data (4 из 4)





Стратегии использования Smart Data.

Пользователи Smart Data

Smart Data для бизнеса:

- Поддержка операционной деятельности (данные для операций);
- Поддержка принятия решений (данные для аналитики);
- Поддержка маркетинговых активностей (аналитика и персонализация);
- Объединение экосистем данных различных предприятий в единые экосистемы по отраслям.

Smart Data для нужд государства:

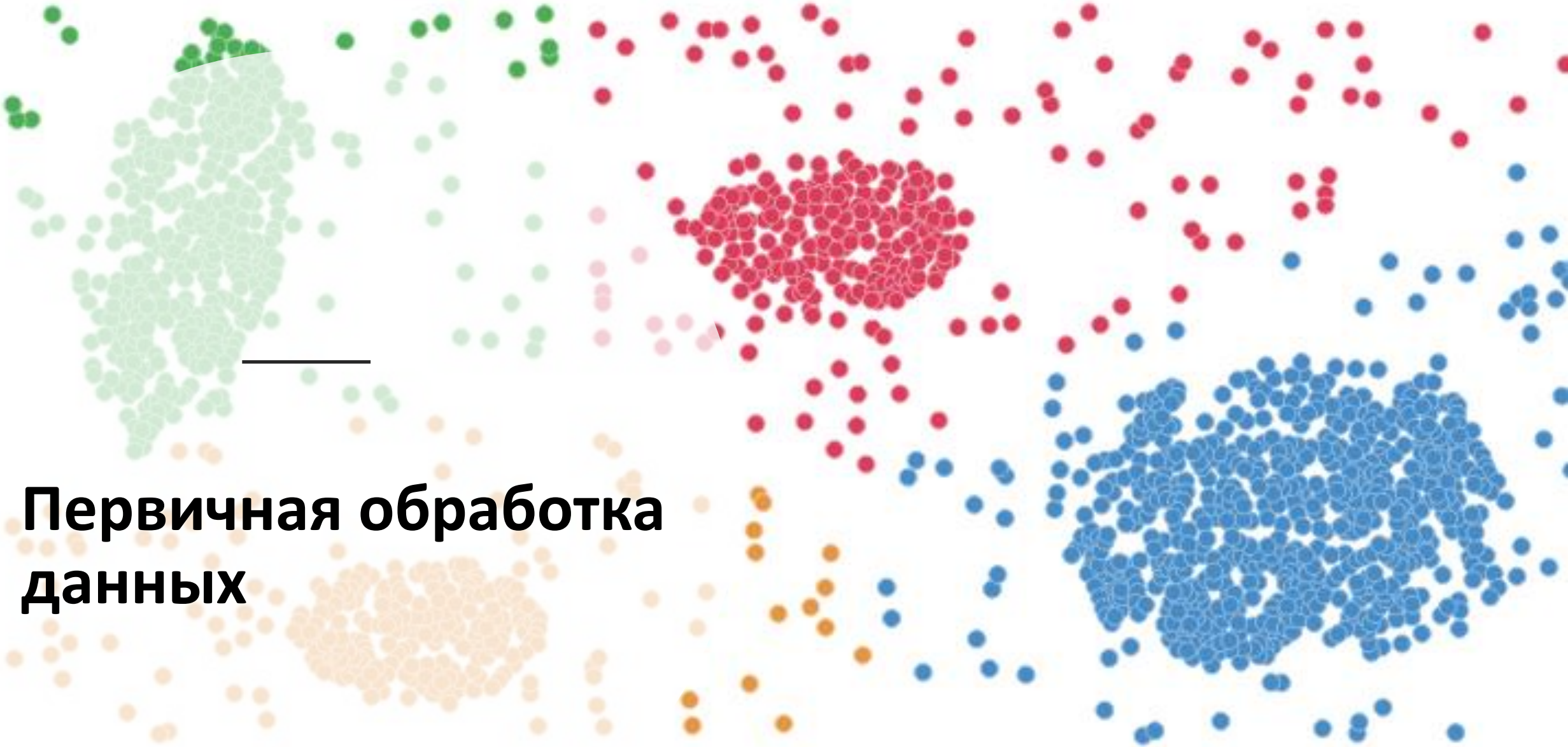
- Развитие транспортной инфраструктуры (телеметрия от транспортных средств);
- Управление малой энергетикой (данные о локальных генераторах, распределение нагрузки);
- Данные в здравоохранении (медицинские исследования, медобслуживание населения, мониторинг состояния здоровья граждан).

Smart Data для общественных объединений (communities).

Основные подходы, используемые в Smart Data

Основные стратегии использования интеллектуальных данных:

- Быстрая первичная обработка данных (например, Event-driven architecture (EDA));
- Быстрая аналитика данных (подготовка первично обработанных данных для передачи далее для принятия решений или поддержки операционной деятельности);
- Быстрое принятие решений (архитектурные решения для быстрых решений)
- Поддержка операционной деятельности / исполнение принятых решений (в общем случае может быть не связано с данными).



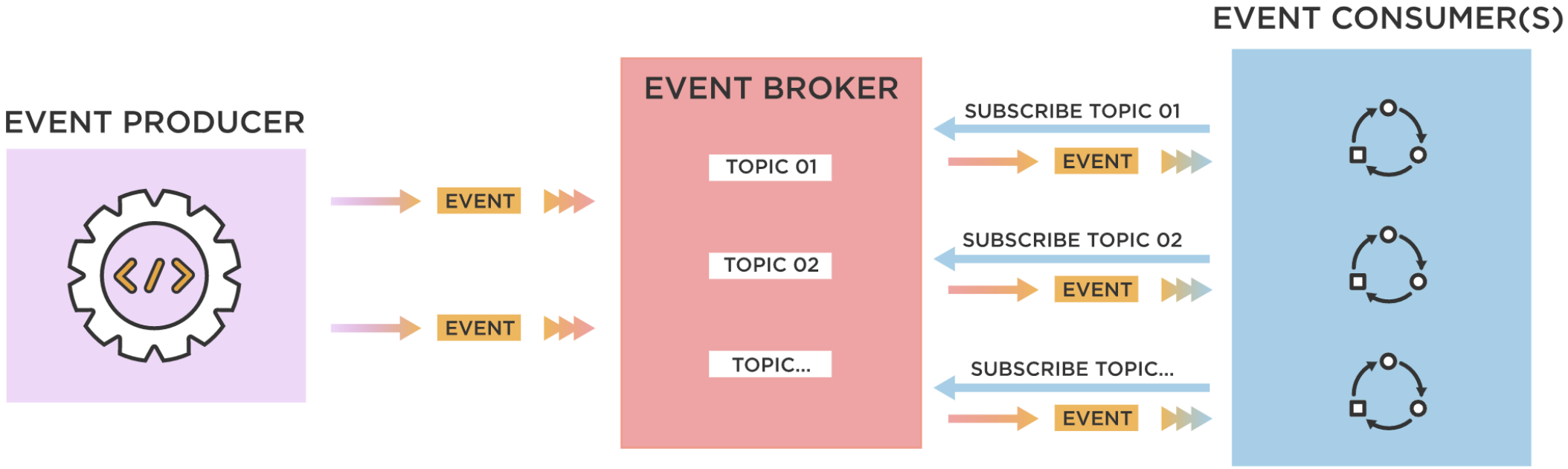
Первичная обработка данных

Методы первичной обработки данных

- Маппинг данных (сопоставление входного потока данных с внутренней информационной структурой);
- Фильтрация данных (удаление данных из потока, которые не подлежат дальнейшему анализу и использованию);
- Удаление дубликатов ассетов данных;
- Простые методы валидации данных (относительно схем и правил), для некоторых случаев возможно дополнение пропущенных фрагментов и исправление ошибок (определяется применяемыми схемами данных).

Event-driven architecture (EDA)

Управляемая событиями архитектура (EDA) - это шаблон проектирования программного обеспечения, который позволяет организации обнаруживать «события» или важные бизнес-операции (такие как транзакция, посещение объекта, брошенная корзина online магазина и пр.), и действовать в соответствии с ними в режиме реального времени или почти в реальном времени. Этот шаблон заменяет традиционную архитектуру «запрос / ответ», в которой службы должны ждать ответа, прежде чем они смогут перейти к следующей задаче.



Event-driven architecture (EDA)

Важные особенности EDA:

- Управляемую событиями архитектуру часто называют «асинхронной» коммуникацией. Это означает, что отправителю и получателю не нужно ждать друг друга, чтобы перейти к следующей задаче. Системы не зависят от этого сообщения.
- При использовании управляемой событиями архитектуры существуют производители событий, которые генерируют и отправляют уведомления о событиях, и может быть один или несколько потребителей события, где получение события запускает логику обработки.
- Традиционно большинство систем работают в парадигме модели, ориентированной на данные, где данные являются источником истины. Переход к архитектуре, управляемой событиями, означает переход от модели, ориентированной на данные, к модели, ориентированной на события. В модели, управляемой событиями, данные по-прежнему важны, но события становятся наиболее важным компонентом.

Event-driven architecture (EDA)

Как работает архитектура, управляемая событиями?

Компоненты событийно-управляемой архитектуры могут включать три части:

- Производитель событий,
- Потребитель событий,
- Брокер событий.

Брокер может быть необязательным, особенно когда есть всего один производитель и один потребитель, которые напрямую взаимодействуют друг с другом, а производитель просто отправляет события потребителю.

Примером может быть производитель, который отправляет события только в базу данных или хранилище данных, чтобы события собирались и сохранялись для анализа. Чаще всего на предприятиях присутствует несколько источников, отправляющих все типы событий, с одним или несколькими потребителями, заинтересованными в некоторых или во всех этих событиях.

Аналитика данных

Data Analysis

Данные сами по себе бесполезны. Они полезны только в том случае, если можно извлечь из них смысл и ценность. Другими словами, важно то, что можно сделать с данными, а не просто то, что они существуют.

Четыре основных вопроса аналитики данных:

- ***Описание:*** *что и когда событие произошло? Как часто оно происходит?*
- ***Объяснение:*** *почему это событие произошло? Каково его влияние на другие рассматриваемые объекты или явления?*
- ***Прогноз:*** *что, скорее всего, произойдет дальше? Что, если бы мы сделали то или это?*
- ***Решение:*** *каков оптимальный ответ или результат? Как это достигается?*

Minelli, M., Chambers, M. and Dhiraj, A. (2013) Big Data, Big Analytics. Wiley, Hoboken, NJ.

Data Analysis

Ответы на эти вопросы получены из четырех основных классов аналитики. Предварительная аналитика при этом подготавливает данные для их обработки в основных классах аналитики:

- ***Предварительная аналитика***
- ***Извлечение данных и распознавание паттернов (Data mining and pattern recognition)***
- ***Визуализация данных и визуальная аналитика.***
- ***Статистический анализ и прогнозирование.***
- ***Прогнозирование, моделирование и оптимизация.***

Minelli, M., Chambers, M. and Dhiraj, A. (2013) Big Data, Big Analytics. Wiley, Hoboken, NJ.

Предварительная аналитика

Основные процессы предварительной аналитики:

- ***Выбор данных:*** *определение подмножества переменных, обладающих наибольшей полезностью, и критериев выборки для этих переменных.*
- ***Предварительная обработка данных:*** *очистка выбранных данных для устранения помех, ошибок или искажений или обработка отсутствующих полей или несоответствия, а также структурирование данных для дальнейшего анализа.*
- ***Сокращение и прогнозирование данных:*** *уменьшение размерности данных за счет преобразования (например, сглаживание, построение атрибутов, агрегирование, нормализация, использование иерархических концепций и статистические методы, такие как регрессионный анализ и анализ основных компонент) для создания эквивалентных, но более эффективных представлений.*
- ***Обогащение данных:*** *объединение выбранных данных с другими данными (например, данными переписи населения, рыночными данными и пр.), чтобы использовать более глубокие знания.*

Miller, H.J. (2010) 'The data avalanche is here. Shouldn't we be digging?', Journal of Regional Science, 50(1): 181–201.

Извлечение данных и распознавание паттернов

Извлечение данных (Data mining) - это процесс извлечения данных и шаблонов из больших наборов данных.

Manyika, J., Chiu, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C. and Hung Byers, A. (2011) Big Data: The Next Frontier for Innovation, Competition, and Productivity. McKinsey Global Institute.

Основной инструмент – методы машинного обучения (см. таблицу).

Table 6.1 Data mining tasks and techniques

Data mining task	Description	Techniques
Segmentation or clustering	Determine a set of implicit groups that describes the data	• Cluster analysis
Classification	Predict the class label that a set of data belongs to based on some training datasets	• Bayesian classification • Decision tree induction • Artificial neural networks • Support vector machine
Association	Find relationships among data objects; predict the value of some attribute based on the value of other attributes	• Association rules • Bayesian networks
Deviations	Find data items that exhibit unusual deviations from expectations	• Cluster analysis • Outlier detection • Evolution analysis
Trends	Lines and curves summarizing the database, often over time	• Regression • Sequence pattern extraction
Generalisations	Compact descriptions of the data	• Summary rules • Attribute-orientated induction

Source: Miller and Han (2009: 7).

Source: Miller and Han (2009: 7).

Визуализация данных и визуальная аналитика

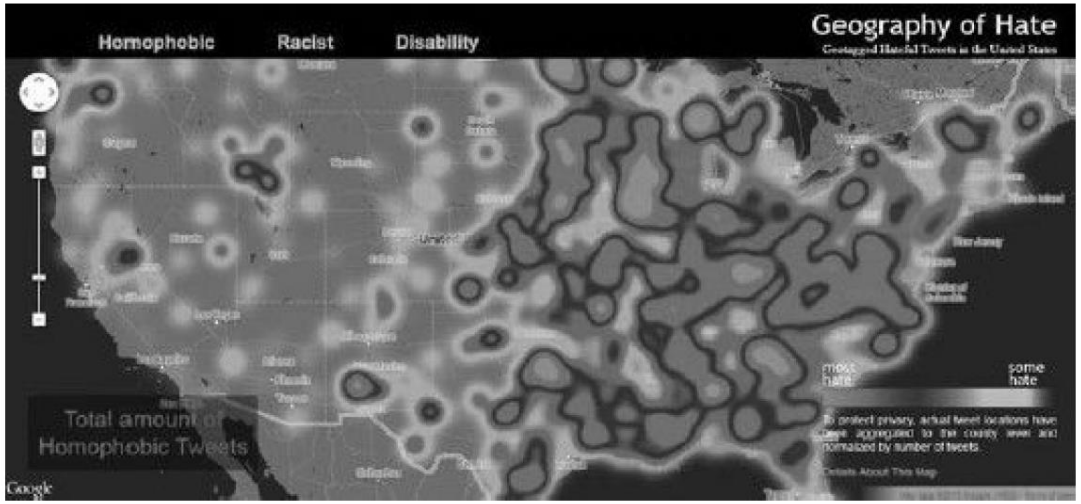


Figure 6.1 The geography of homophobic tweets in the United States

Source: http://users.humboldt.edu/mstephens/hate/hate_map.html#

Figure 6.2 Real-time flight locations

Source: http://www.flightradar24.com/

Figure 6.3 CASA's London City Dashboard

Source: http://citydashboard.org/london/

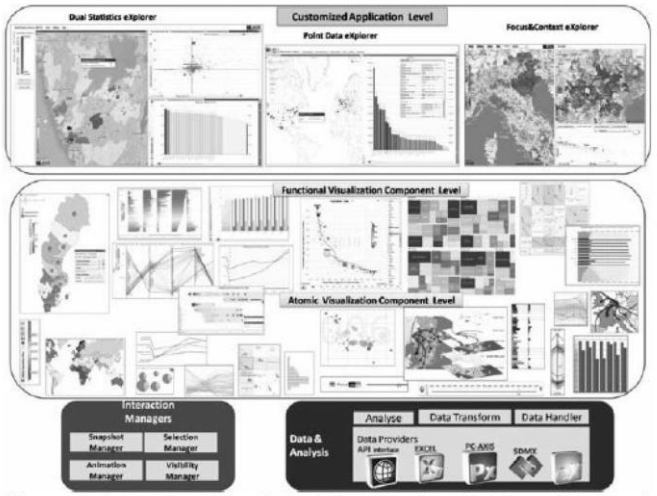


Figure 6.4 Geovisual Analytics Visualization (GAV) toolkit developed by the National Center for Visual Analytics, Linköping University

Статистический анализ

Обзор статистических методов, используемых для анализа больших данных:

- ***Описательная статистика*** *(подробно описывает характеристики и распределение данных, а также уровни ошибок и неопределенности). Включает в себя:*
  - *Анализ временных рядов (параметры, изменяющиеся во времени);*
  - *Теория графов (математическое описание организационных и сетевых структур);*
  - *Пространственная статистика (описывает геометрию и паттерны пространственной кластеризации, дисперсии и диффузии).*
- ***Статистика вывода (Inferential statistics)*** *(стремится объяснить, а не просто описать закономерности и взаимосвязи, которые могут существовать в рамках набор данных, а также для проверки силы и значимости ассоциаций между переменными). Включает в себя:*
  - *Параметрическая статистика:*
  - *Непараметрическая статистика;*
  - *Вероятностная статистика.*

Kitchin, R. (2014). The data revolution: Big data, open data, data infrastructures & their consequences. SAGE Publications Ltd https://www.doi.org/10.4135/9781473909472

Прогнозирование, моделирование и оптимизация

1. ***Прогнозирование.*** *Предлагается ансамблевый подход для решения задач прогнозирования. При таком подходе строится множество прогнозных моделей, основанные на разных методах (например, временные ряды, поведенческие модели и пр.), по каждой из моделей строится прогноз и результаты объединяются.*

Franks, B. (2012) Taming the Big Data Tidal Wave: Finding Opportunities in Huge Data Streams with Advanced Analytics. Wiley, Hoboken, NJ.

2. ***Моделирование*** *- это построение моделей, которые стремятся моделировать реальные процессы и системы. Цель состоит в том, чтобы определить как функционирует система и как она может вести себя в различных сценариях, а также статистическая оценка их эффективность с целью повышения эффективности и результативности.*

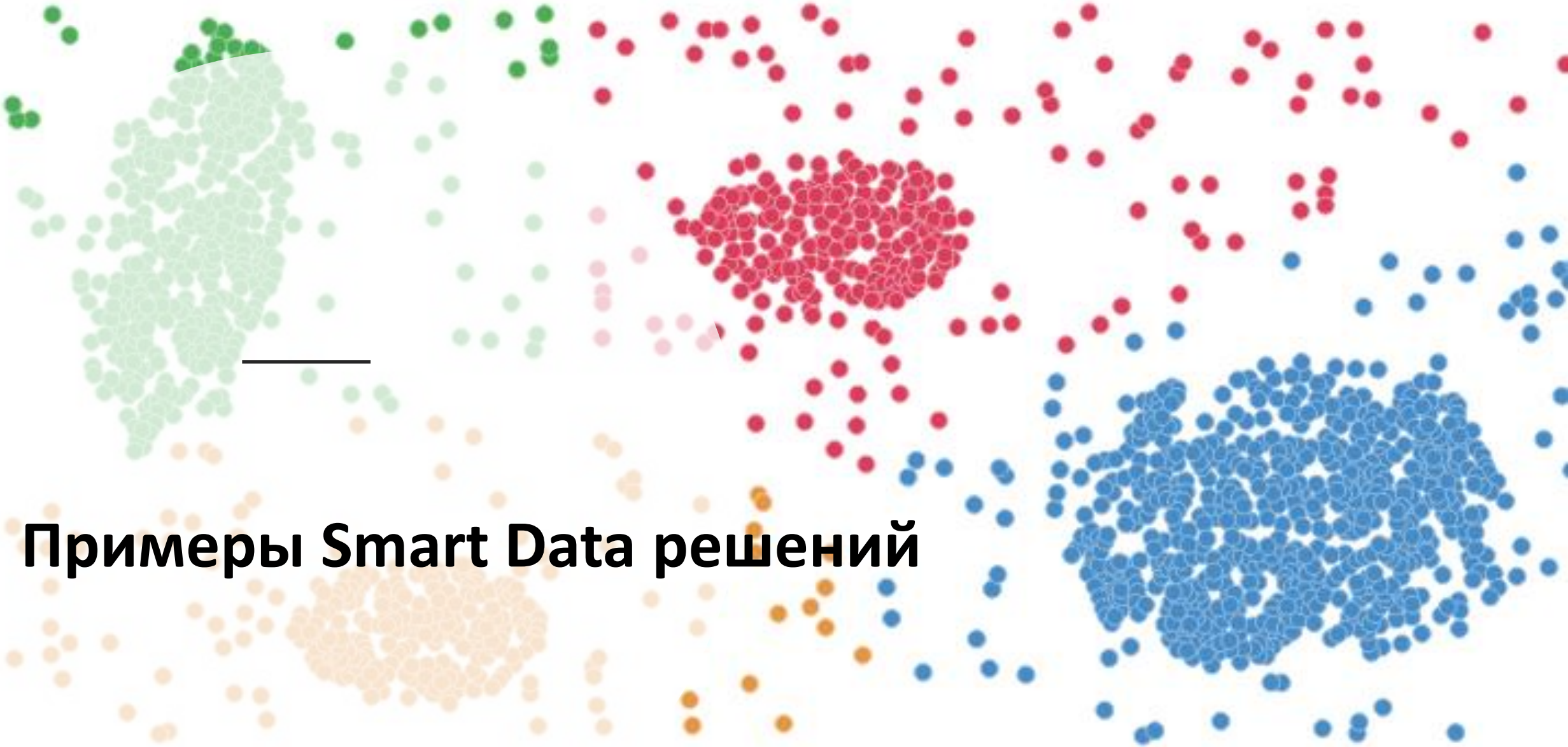
Robinson, S. (2003) Simulation: The Practice of Model Development and Use. John Wiley & Sons, Chichester.

Прогнозирование, моделирование и оптимизация

3. ***Оптимизация*** *связана с определением оптимального курса действий для повышения производительности (как правило, снижение затрат или увеличение объема производства/оборота). Такой курс можно вычислить, используя и оценивая прогностические и имитационные модели, либо могут быть разработаны другие виды алгоритмов или статистических тестов.*

   Например, генетические алгоритмы - особый вид машинного обучения, который используют идеи из естественного отбора, такие как наследование, мутация, отбор и скрещивание, для развития и эволюции возможных решений проблемы.

   Mitchell, M. (1996) An Introduction to Genetic Algorithms. MIT Press, Cambridge, MA.



Примеры Smart Data решений

Примеры Smart Data решений

1. *PRO-OPT – Big Data Production Optimisation in Smart Ecosystems.*
   http://www.pro-opt.org/
2. *SIDAP – Scalable Integration Concept for Data Aggregation, Analysis and Preparation of Big Data Volumes in Process Manufacturing.*
   http://www.sidap.de/
3. iTESA – intelligent Traveller Early Situation Awareness .
   http://www.smart-data-itesa.de/
4. *sd-kama – Smart Data Disaster Management.*
   http://www.sd-kama.de/de/smart_data_disaster_management/
5. *SD4M – Smart Data for Mobility.*
   http://www.sd4m.net/
6. InnOPlan – Innovative, Data-driven Efficiency of Surgery-related Process Landscapes.
   https://innoplan.uni-hohenheim.de/

Благодарю за внимание