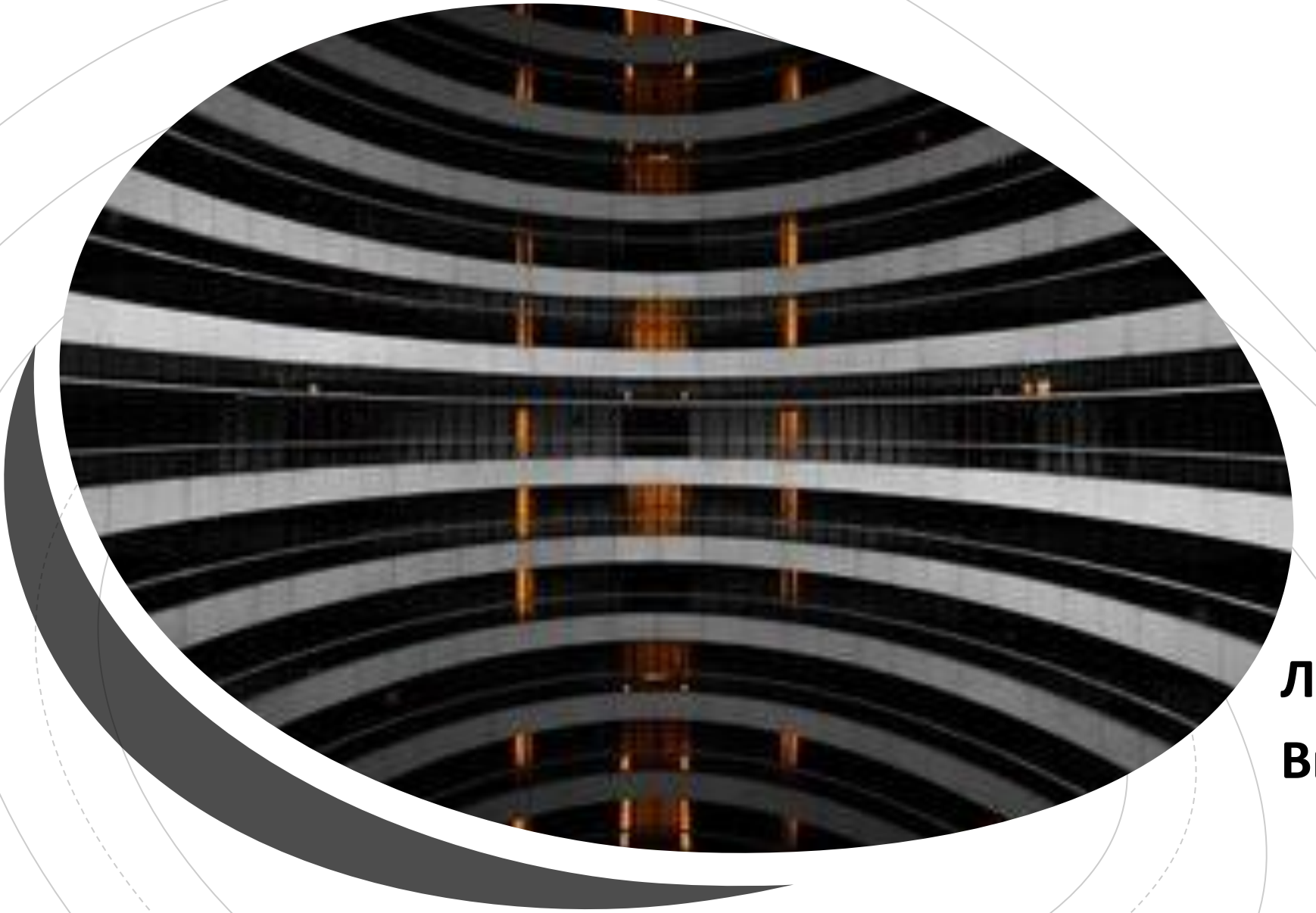


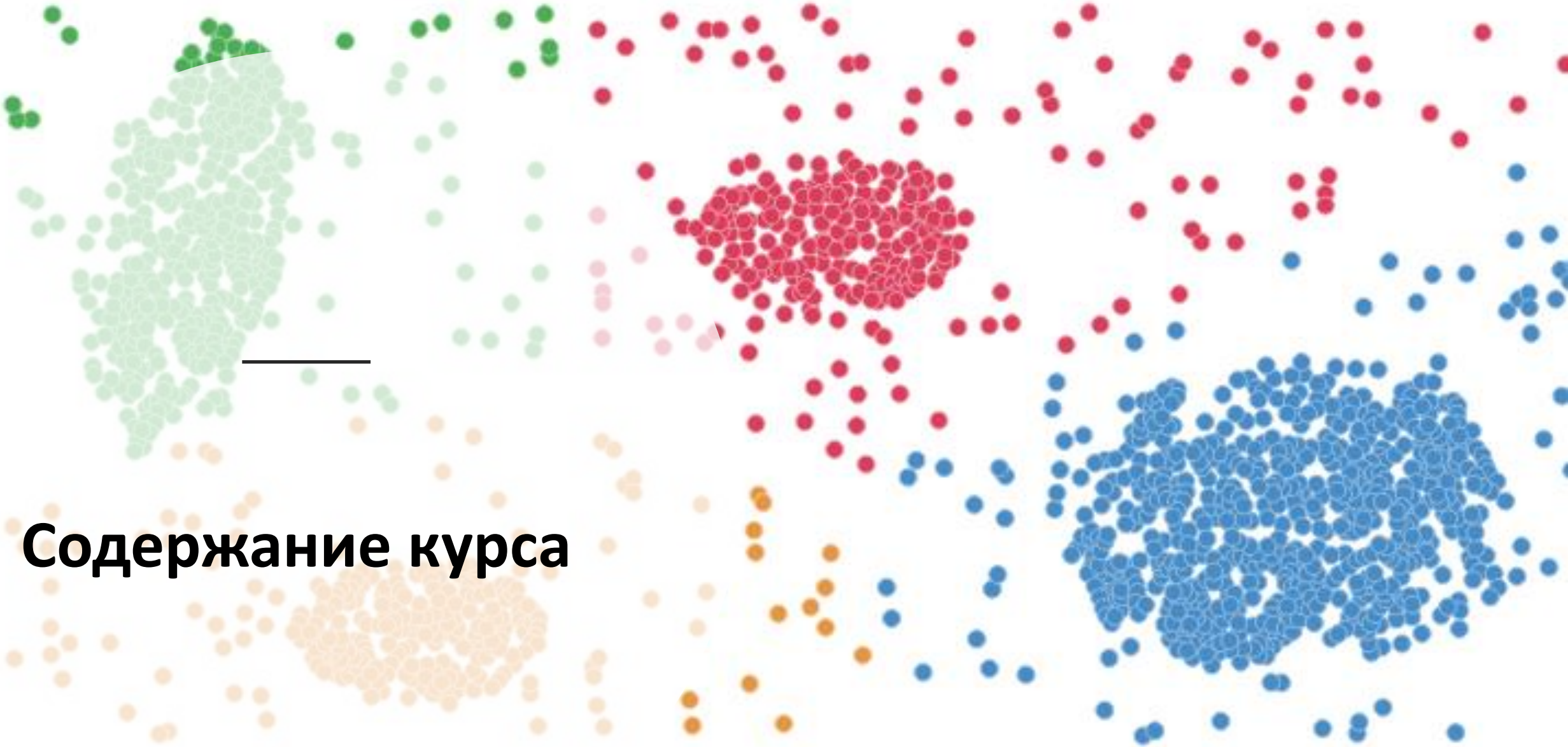
Smart data

Лекция #1

Введение

И.А. Куликов

i.a.kulikov@gmail.com



Содержание курса

Содержание курса Smart data (Интеллектуальные данные)

- **Лекции:**

- Введение
- Большие данные и хранилища данных
- Графы знаний
- Основы машинного обучения
- Традиционные БД
- Семантические данные и Semantic Metaminig

- **Лабораторные работы / практические занятия**

- **Исследовательская работа**

Основные определения

Данные

- Интерпретируемое формализованным способом представление информации, пригодное для коммуникации, интерпретации или обработки. [ИСО/МЭК 2382-1] [ГОСТ Р 52292-2004]

- Информация, представленная в виде, пригодном для обработки автоматическими средствами при возможном участии человека [ГОСТ 15971-90] [ГОСТ Р 50304-92] [ОСТ 45.127-99]

Данные

- Данные — зарегистрированная информация: представление фактов, понятий или инструкций в форме, приемлемой для общения, интерпретации, или обработки человеком или с помощью автоматических средств (ISO/IEC/IEEE 24765-2010).

В информатике и информационных технологиях:

- Данные — поддающееся многократной интерпретации представление информации в формализованном виде, пригодном для передачи, связи или обработки (ISO/IEC 2382:2015).

- Данные — формы представления информации, с которыми имеют дело информационные системы и их пользователи (ISO/IEC 10746-2:1996)

Большие данные (Big data)

По данным Gartner -
«Большие данные» - это объемные, быстро доступные и разнообразные информационные ресурсы, которые требуют рентабельных, инновационных форм обработки информации для лучшего понимания и принятия решений.

Чаще всего о больших данных говорят как о таких данных, *размер которых превышает способность типовых СУБД собирать, хранить, управлять и анализировать их. А также анализ и обработки которых может иметь большую ценность для бизнеса*.

Интеллектуальные данные (Smart data)

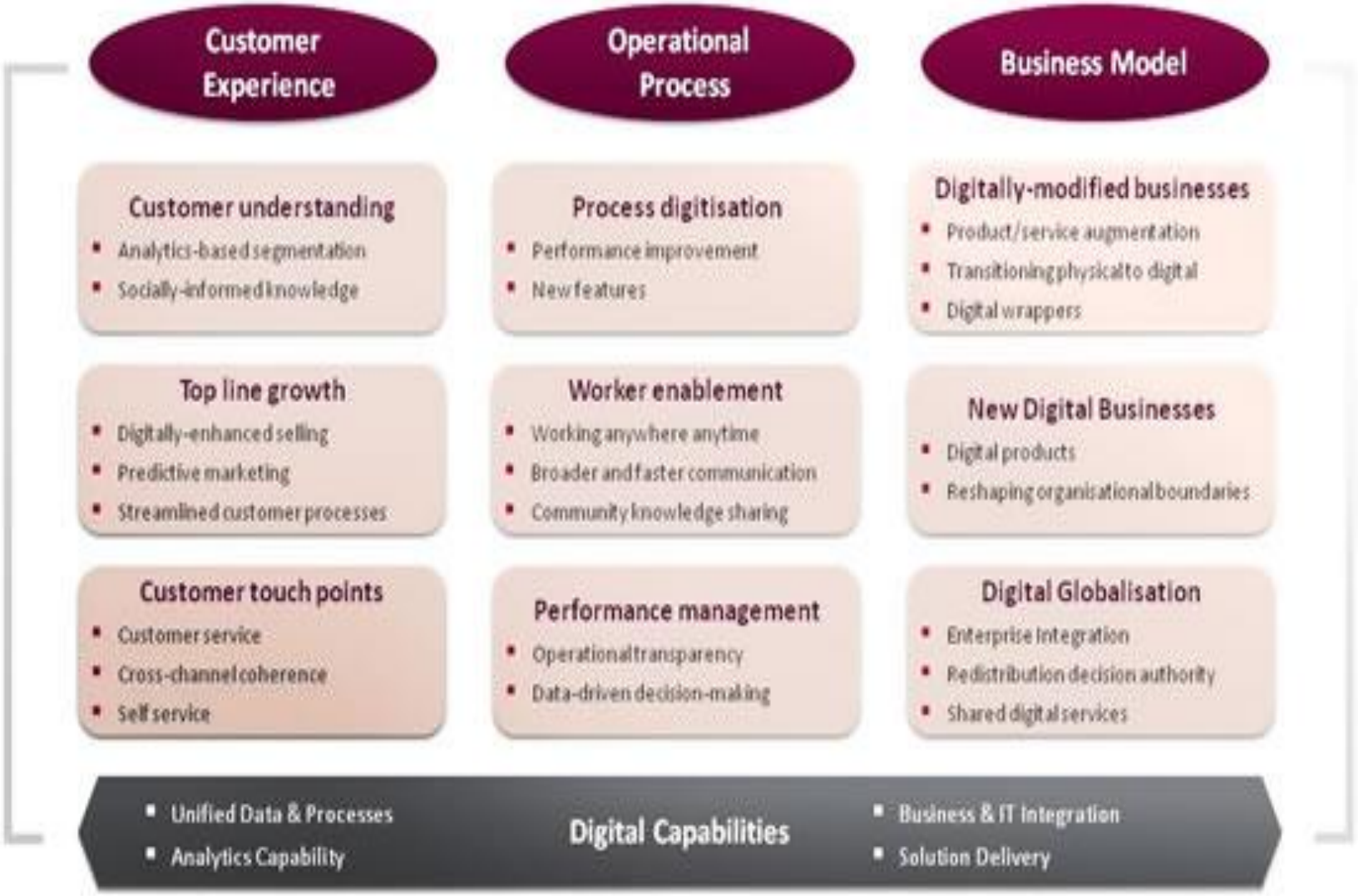
Интеллектуальные данные - это данные, которые предварительно обработаны таким образом, чтобы они могли быть использованы для решения задач непосредственно в точки их сбора перед отправкой на платформу аналитики для дальнейшей консолидации и анализа.

Важность данных для процессов цифровой трансформации

Области цифровой трансформации

- Бизнес деятельность / функции
- Бизнес-процессы
- Бизнес-модели
- Бизнес-экосистемы
- Управление бизнес-активами
- Организационная культура
- Экосистемные и партнерские модели
- Подходы к клиентам, работникам и партнерам.

Цифровая трансформация происходит повсюду



Цифровая трансформация происходит повсюду

В приведенном ниже списке от McKinsey показаны те аспекты, в которых может сыграть роль цифровая трансформация:

- (Цифровой) клиентский опыт.
- Инновации в продуктах и услугах.
- Распределение, маркетинг и продажи.
- Цифровое исполнение, оптимизация рисков, усиленный корпоративный контроль и т. д.

Цифровая трансформация и ключевая роль данных

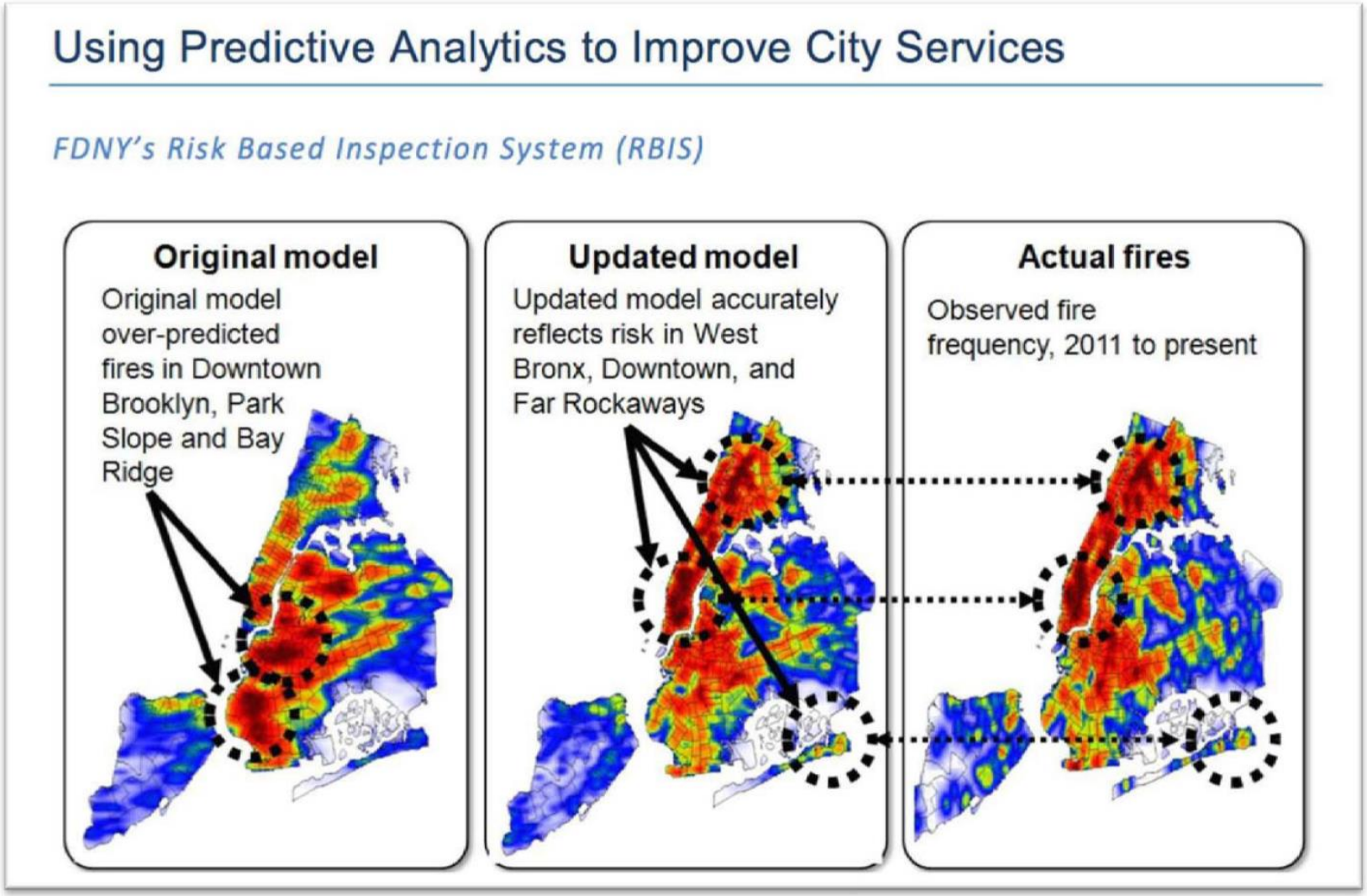
Цифровая трансформация и интеллектуальные данные (Smart data)

Новые элементы уравнения и преобразования информации :

- Интеллект (например, искусственный интеллект).
- Скорость (скорость - это качество обслуживания клиентов и даже конкурентное преимущество).
- Целостный подход к безопасности (с информацией и данными в качестве активов).
- Необходимость оцифровывать и фиксировать бумажные данные (цифровое преобразование требует оцифровки и, следовательно, сканирования) ближе к источнику, владельцу и процессу, чтобы перейти на безбумажный документооборот.
- Повышенное внимание к точности, качеству и результатам.

Большие данные и конфиденциальность

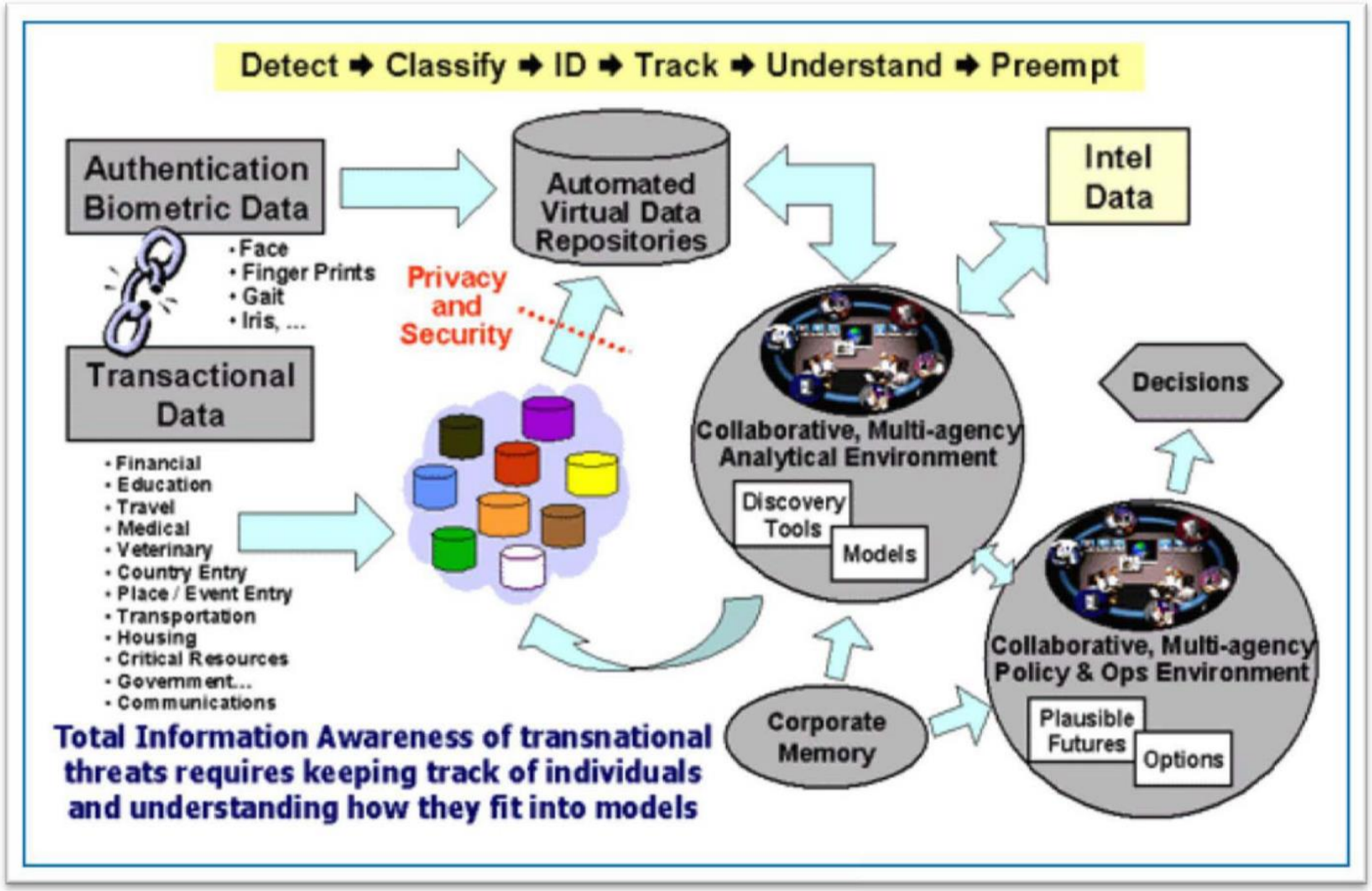
Данные для безопасности



New York City

Courtesy of Mayor's Office of Data Analytics. Used with permission.

Данные для безопасности



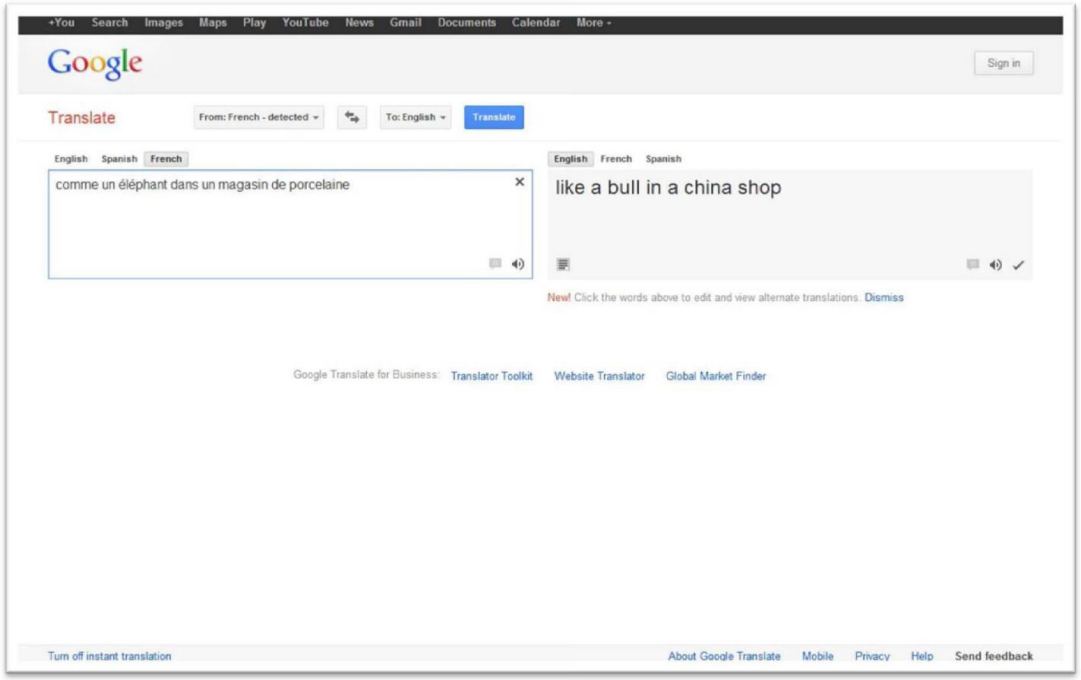
Total Information Awareness program

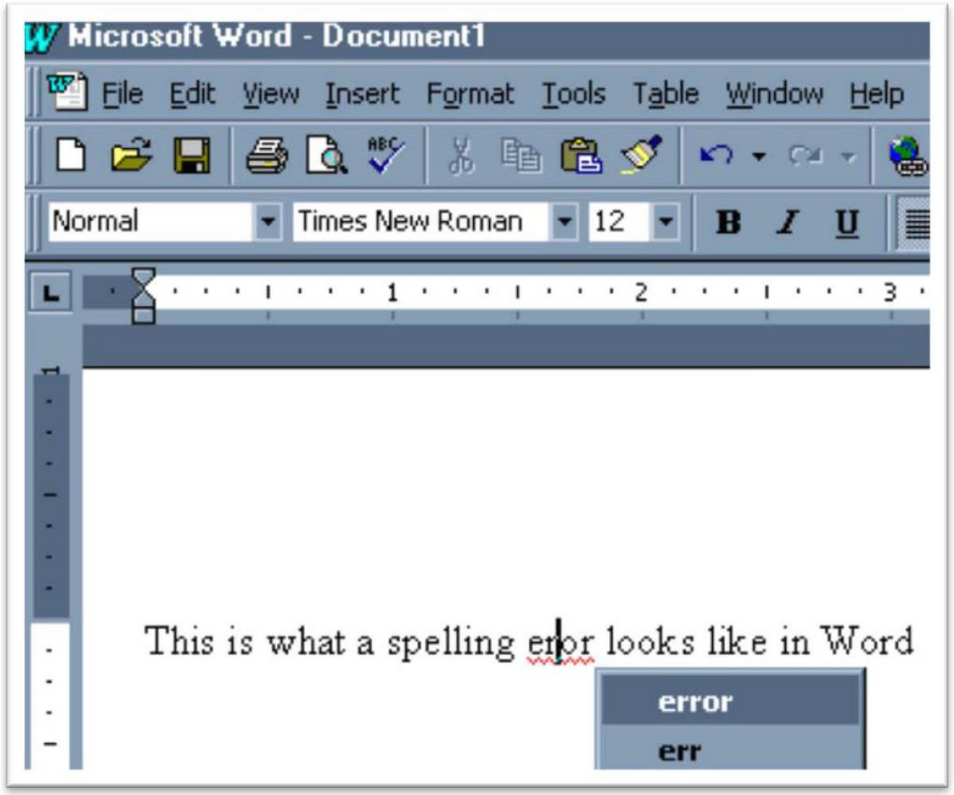
Данные для безопасности

Future Attribute Screening Technology

This image is in the public domain.

Данные для сервисов





Dataveillance – наблюдение за данными

Определение: «систематический мониторинг людей или групп, с помощью систем обработки персональных данных, чтобы регулировать или управлять их поведением ». (Degli Esposti, 2014)

Приложения:

- ❑ Найм и удержание персонала
- ❑ Лояльность потребителей
- ❑ Эффективность цепочки поставок
- ❑ Безопасность и предотвращение рисков

Dataveillance – наблюдение за данными

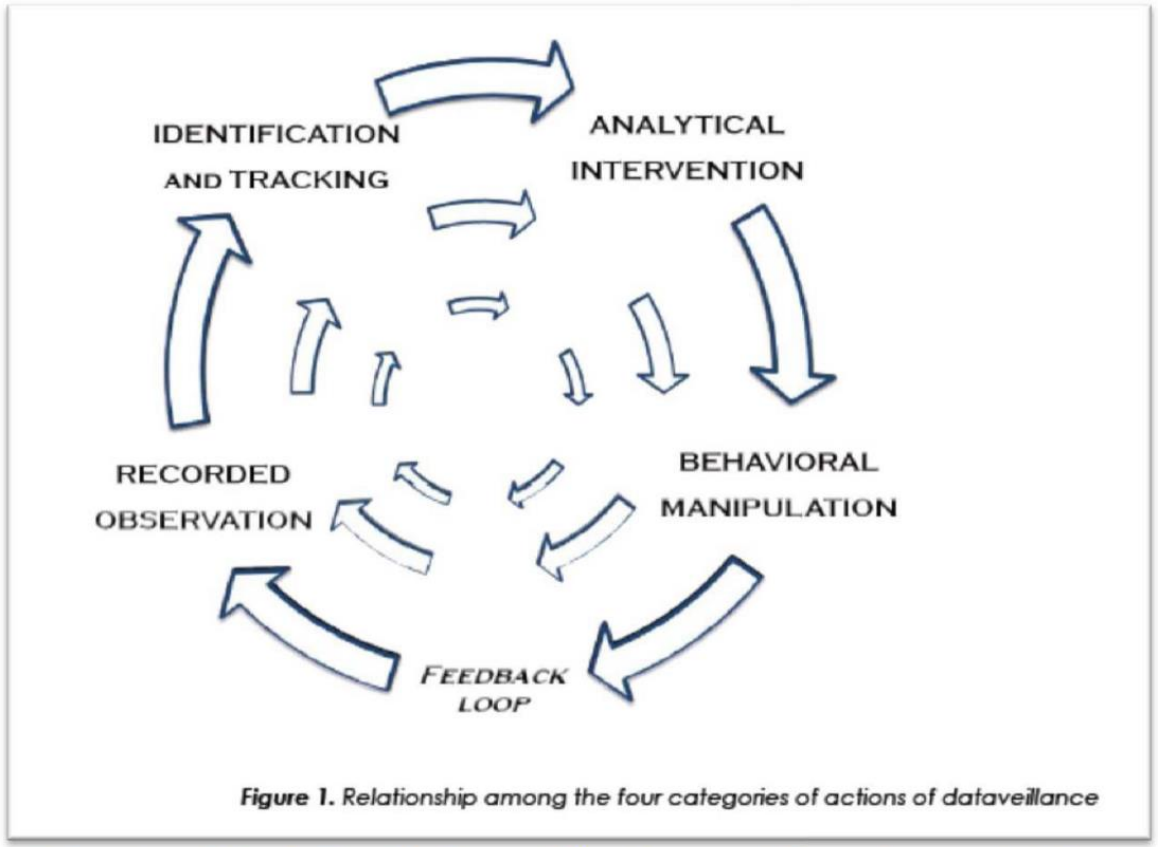


Figure 1. Relationship among the four categories of actions of dataveillance

Основные проблемы:

- ❑ Неравные отношения между ИС и клиентами
- ❑ Это происходит без нашего ведома

(Ref. Degli Esposti, 2014)

Гражданские свободы

Использование больших данных может приводить к следующим негативным последствиям для граждан:

- ❑ Платить больше за медицинское страхование
- ❑ Исключены из медицинского страхования
- ❑ Получить отказ в ссуде
- ❑ Подвергаться профилированию или дополнительной проверке
- ❑ Содержаться в тюрьме
- ❑ Быть арестованным до совершения преступления
- ❑ Быть признанным виновным по соучастию

Гражданские свободы

Использование больших данных может приводить к следующим негативным последствиям для граждан:

- ❑ Платить больше за медицинское страхование
- ❑ Исключены из медицинского страхования
- ❑ Получить отказ в ссуде
- ❑ Подвергаться профилированию или дополнительной проверке
- ❑ Содержаться в тюрьме
- ❑ Быть арестованным до совершения преступления
- ❑ Быть признанным виновным по соучастию

Гражданские свободы

Обвинение человека в возможном поведении способно:

- ❑ Отрицать саму основу справедливости(презумпция невиновности и ответственность людей)
- ❑ Отрицать человеческую волю и свободу воли
- ❑ Освобождают людей от любой ответственности

Должны ли мы согласиться ограничивать нашу личную свободу и полагаться на аналитику больших данных для общего блага?

Может ли аналитика больших данных помочь нам избежать гендерной дискриминации и расового профилирования?

Конфиденциальность и гражданские свободы

Действующие стратегии:

- ❑ Уведомление и индивидуальное согласие
- ❑ Не принимать участия
- ❑ Анонимность
- ❑ Нейтрализация нарушителей конфиденциальности

Конфиденциальность и гражданские свободы

Возможные решения:

- ❑ Переход от «конфиденциальности по согласию» к «конфиденциальности по подотчетность»
- ❑ Определение новых правил (на уровне правительств или рынков)
- ❑ Новое определение понятия справедливости для сохранения свободы воли человека
- ❑ Открытие доступа к наборам данных и алгоритмам (прозрачность)
- ❑ Поручение беспристрастным экспертам сертифицировать алгоритмы и их реализацию
- ❑ Размытие наборов данных (дифференциальная конфиденциальность)
- ❑ Определение того, как люди могут опровергнуть предсказание о них

Конфиденциальность и гражданские свободы

Возможные решения (Bollier, 2010):

- ❑ Стирание данных по истечении заданного периода времени
- ❑ Предоставление людям права владеть своими данными
- ❑ Ограничение сбора данных до необходимого уровня
- ❑ Поощрение самодисциплины путем введения новых социальных норм
- ❑ Обыгрывать систему, добровольно изменяя наше поведение
- ❑ Проявлять инициативу

Методологически значимо:

Смена парадигмы - большие данные требуют от исследователей:

- ❑ Справляться с изобилием и принимать беспорядок в данных
- ❑ Проводить исследования, не задавая точных вопросов
- ❑ Получать представление о данных (индукция) вместо тестирования теории путем анализа данных (дедукция)
- ❑ Использовать сотни алгоритмов вместо выбора одного
- ❑ Перестать зацикливаться на причинно-следственных связях и довольствуйтесь корреляцией

Методологически значимо:

Преимущества анализа больших данных по сравнению с традиционными научными методами:

- Снижает риск ошибки и предвзятости, связанных с отбором выборок
- Дает возможность создавать более динамичные и сложные модели
- Помогает найти корреляции, которые никакая теория не может определить(но может привести к серьезным ошибкам)
- Позволяет рассматривать данные, когда-то воспринимаемые как «шум»

Методологически значимо:

ДВА ПОДХОДА:

Эмпирический (бизнес):

- Большие данные могут говорить сами за себя, не нуждаясь в теориях, моделях или гипотезах (ошибочная).
- Аналитика больших данных свободна от человеческих предубеждений. Их можно интерпретировать кто угодно и их значения выходят за рамки контекстов (ошибочно).

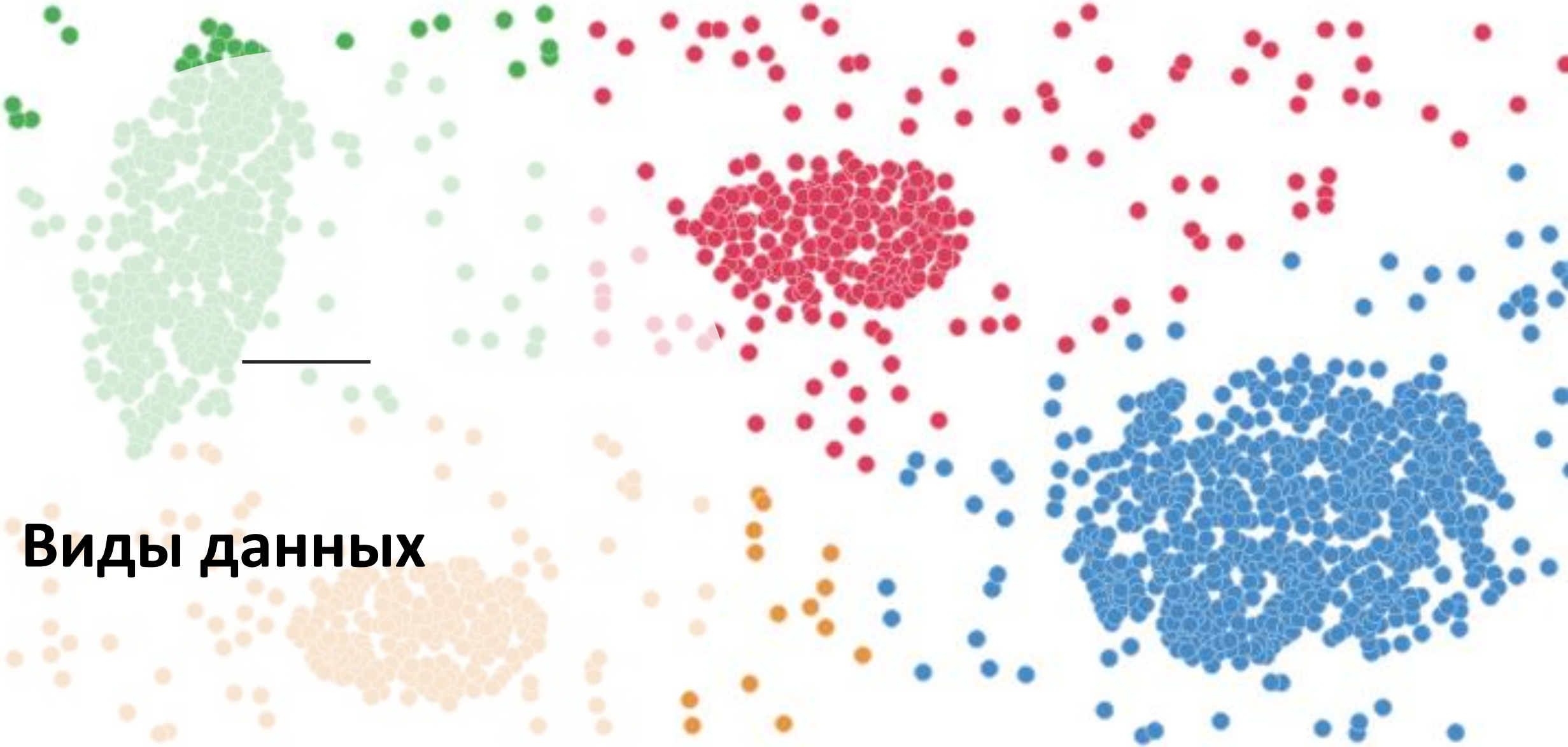
Наука, управляемая данными (академия):

- Использование существующих теорий и концепций для анализа наборов данных.
- Использование больших данных для выявления правильных вопросов и формулирования гипотез.
- Использование наборов данных для ответа на вопросы, для которых эти наборы данных не были предназначены (задача).

Методологически значимо:

Альтернативный путь, предложенный Китчиным (2014):

- Опираясь на критическую теорию, определить, как исследования проводятся и как результаты интерпретируется.
- Признание того, что исследования никогда не идеологически нейтральны и объективны.
- Дополнение исследований больших данных исследованиями небольших данных.



Виды данных

Виды данных.

Данные различаются по:

- **Форме** (качественной или количественной)
- **Структуре** (структурированная, полу-структурированная или неструктурированная)
- **Методу получения** (захваченные, производные, исчерпывающие, переходные)
- **Виду источника данных** (первичный, вторичный, третичный)
- **Типу** (индексный, атрибут, метаданные).

Виды данных.

Альтернативные способы классификации данных:

- Данные с технической точки зрения
- Данные с этической точки зрения
- Данные с политической и экономической точек зрения
- Данные во времени и пространстве
- Данные с философской точки зрения

Пирамида знаний

Традиционная пирамида знаний

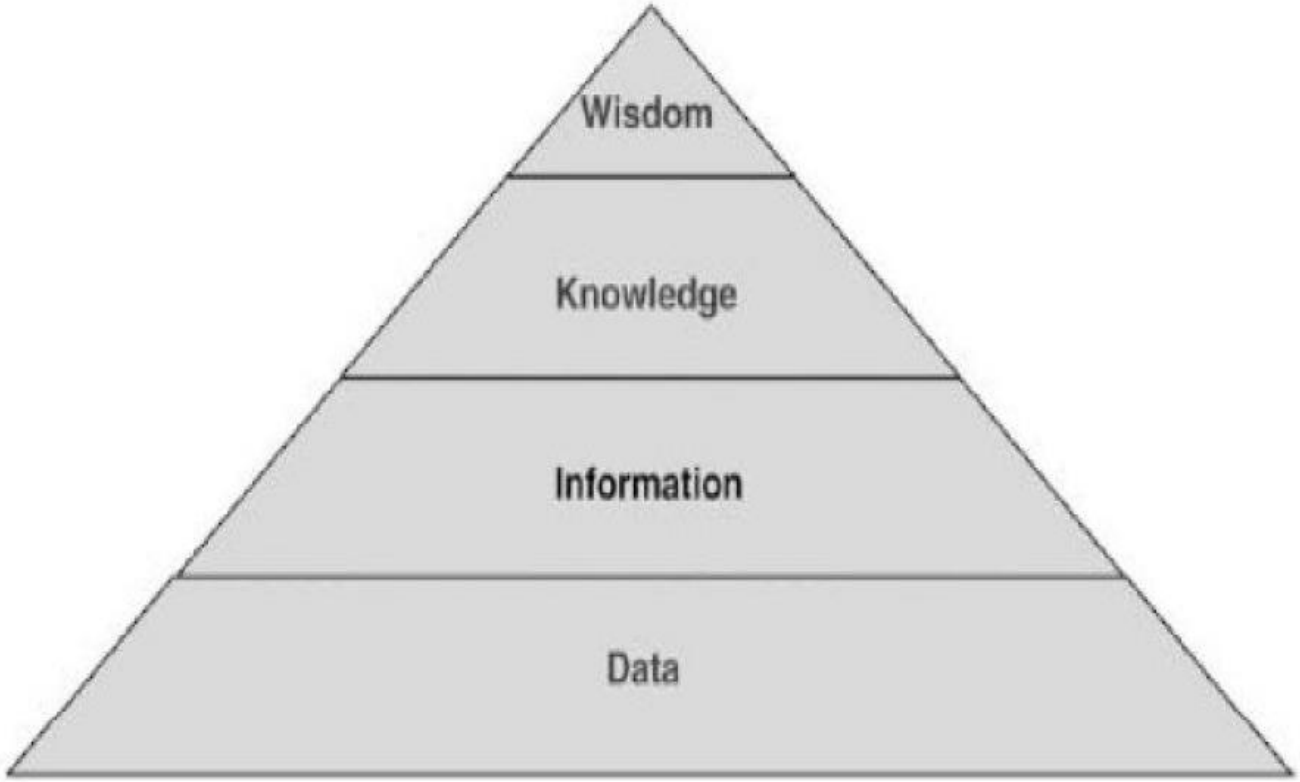


Figure 1, The Knowledge Pyramid [1]

Пересмотренная пирамида знаний

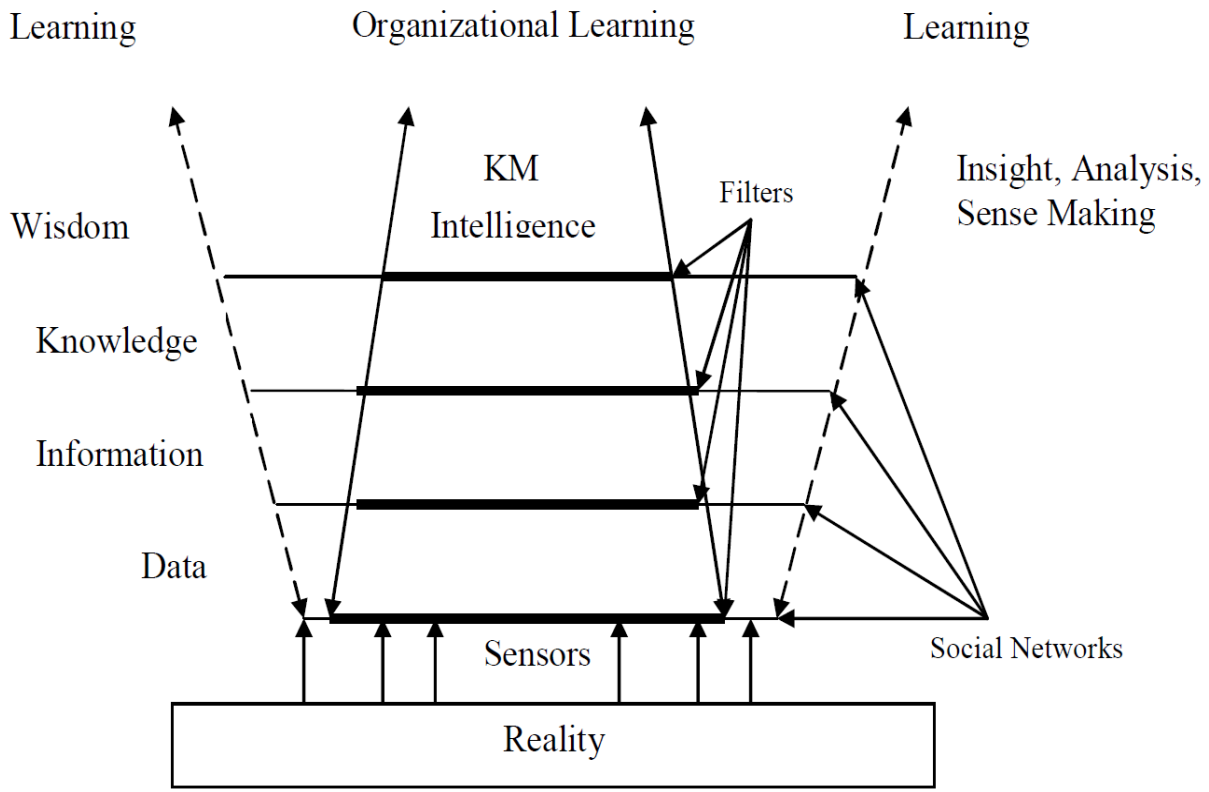


Figure 2, Revised Knowledge Pyramid

Благодарю за внимание