Модели данных

B1_Расширенное качество данных

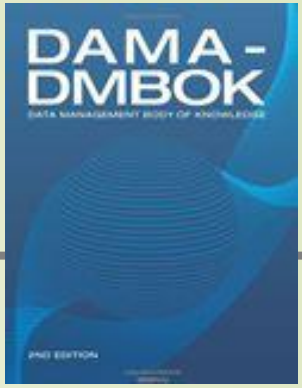
Московский государственный технический университет
имени Н.Э. Баумана

Факультет ИБМ

Июль 2024 года

Москва

Что такое управление данными?



История DQ:
Ввод данных
Сбор данных
Базы данных
Витрины для BI
Хранилища данных
Аналитические модели

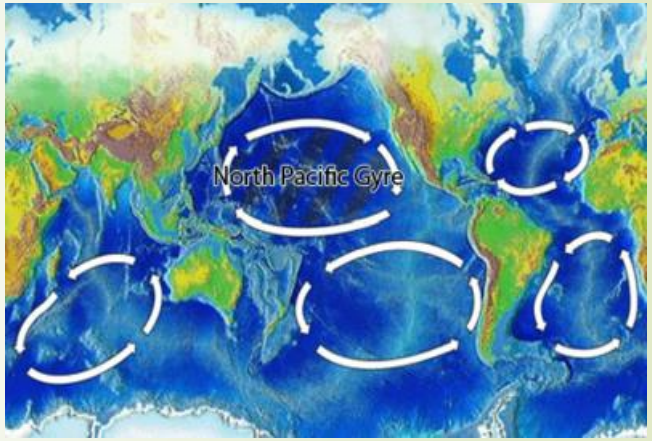
- **Управление данными** рассматривает информационные ресурсы как корпоративные активы для извлечения из них ценности с точки зрения бизнеса
- **Меры ценности информационного актива**
  - критичность
  - важность
  - *качество*
  - стоимость
- **Качество данных** – соответствие данных согласованным требованиям к данным с точки зрения их применения

Принципы, важные для качества данных

- **Воспроизводство мусора GIGO** (англ.: Garbage In – Garbage Out)

  «Грязные» данные влияют
  - на исполнение операций (операционные риски)
  - на качество аналитических моделей и принимаемых решений (управленческие риски)
  - на качество публикаций в Интернет и в печати (репутационные риски)

Тихоокеанский «мусороворот»



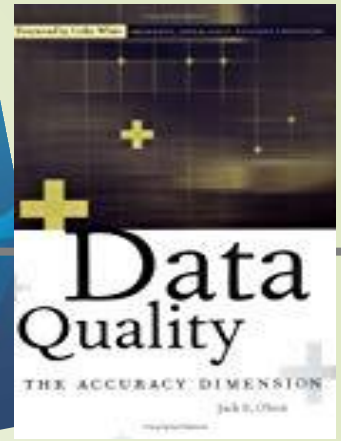
- **Однозначность элементов данных**

  Для представления (хранения) значений каждой величины должен отводится отдельный элемент данных.
  Исключение: строчные композиты (например: ФИО, адреса и т.п.)

- **Историчность данных**
  - явная привязка фактов ко времени свершения (наблюдения)
  - указание периода актуальности для записей справочников и реестров
  - данные имеют определённую историческую глубину

- **Неопределённость при оценке качества данных**
  - либо получаем достоверные, но не совсем актуальные данные
  - либо получаем актуальные данные, но не совсем достоверные

Управление качеством данных (Data Quality Management)



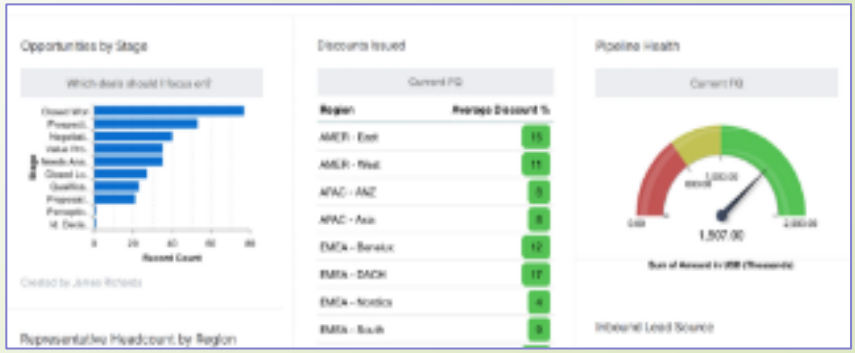
- **Управление качеством данных –**
  обеспечение соответствия состояния данных согласованным требованиям пользователей данных
- **Непрерывный и распределённый процесс на всём ЖЦ данных**
  - подготовка данных
  - сбор данных
  - преобразование и загрузка данных в ХД
  - преобразование и загрузка данных в витрину/ песочницу
  - обработка данных
- **Развитие не новой для нас темы**

  Не просто проверка данных при их сборе и загрузке, но также:
  - *профилирование* – начальное ознакомление с источниками данных
  - *измерение качества данных* путём контроля данных и расчёта метрик
  - *наблюдение и реагирование* на превышение установленных уровней ошибок
  - *очистка*: корректировка и преобразование данных
  - *раскрытие сведений*: публикация отчётов, визуализация на информационных панелях, доступ к данным с отметками об ошибках

Мусоро-сжигательный завод Вены

Принципы управления качеством данных

- **Прозрачная и понятная модель показателей качества данных**
  для однозначной оценки состояния качества данных на основе контроля данных и расчёта метрик показателей согласно требованиям пользователей
- **Сбалансированность требований к качеству данных**
  учёт рисков использования «грязных» данных, затрат на их выявление и исправление, а также полученного эффекта
- **Максимальная автоматизация**
  управление синтаксическим качеством данных наиболее автоматизируемо
- **Информированность о качестве данных**
  предоставление участникам (прежде всего, пользователям) сведений о качестве данных, которые нужно учитывать при работе с данными
- **Институт владельцев и кураторов данных**
  выделение ролей для экспертизы данных предметной области, в том числе с точки зрения их качества



Расширение свойств качества данных

- Традиционное понимание качества данных
- Контроль данных автоматизирован

- Расширение традиционного качества данных
- Не удаётся автоматизировать контроль всех свойств

Характеристики доступности данных

Данные перечислены в каталоге данных или на центральном портале для доступа к наборам данных?

Возможен контекстный поиск данных по метаданным?

Данные консолидированы и интегрированы?
Регулируется доступ, назначены владельцы данных?

Возможно подключение, просмотр и перемещение данных в песочницу?

Можно обнаружить, маскировать или блокировать чувствительные данные?

Есть описания потоков и компонентов данных в виде бизнес-метаданных?

Описания ясны и понятны для бизнес-потребителей данных?

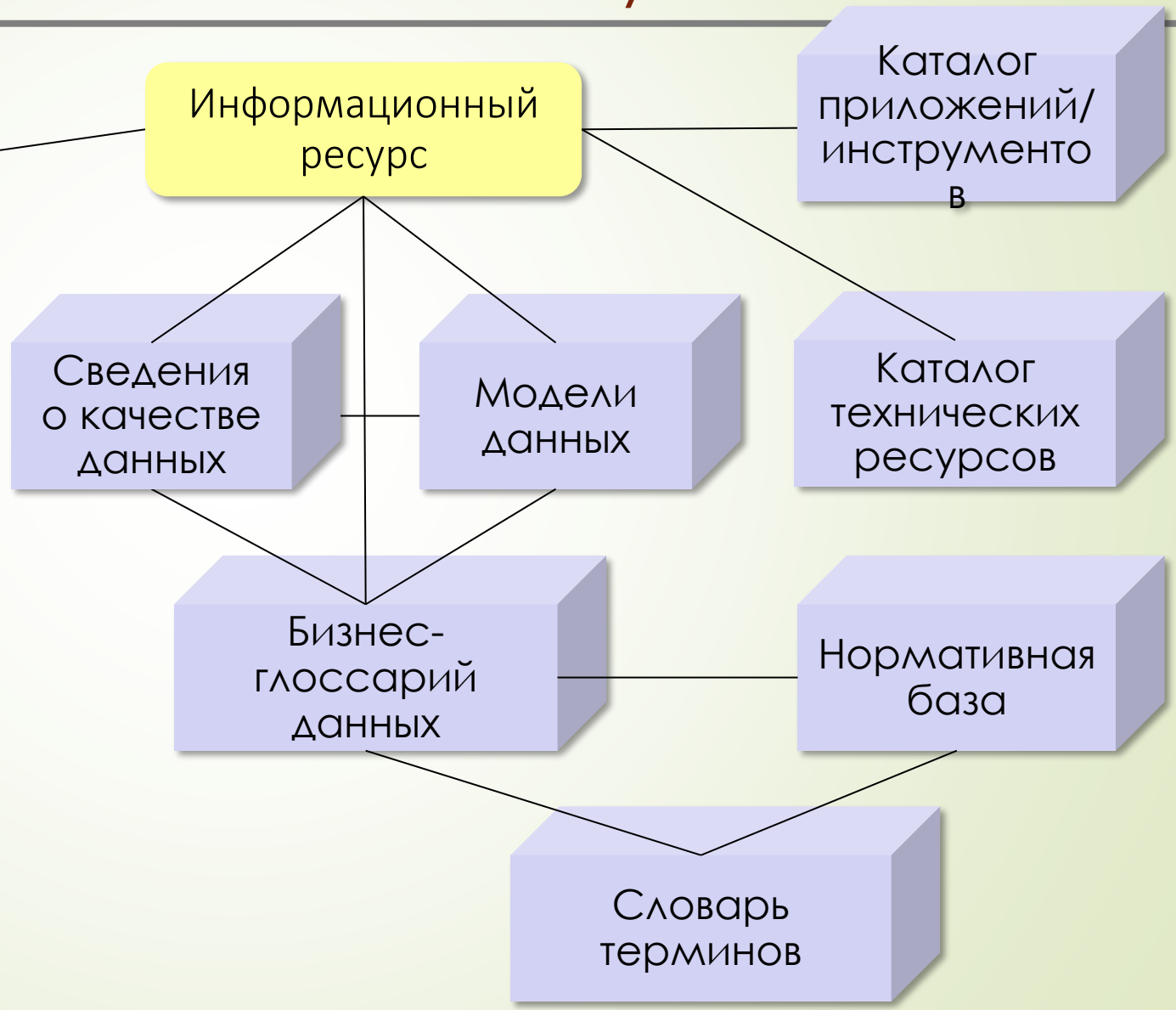
Имеются модели данных и описания форматов и интерфейсов?

Понятны модели, форматы и API специалистам ИТ и исследователям?

Легко данные могут быть обработаны бизнес-пользователями?

8 Каталог данных для оценки доступности данных

- Назначение данных
- Описание данных
- Источники данных
- Владелец/ кураторы
- Потребители/ поставщики
- Бизнес-процессы
- Права доступа
- Параметры подключения
- Политики, процедуры, стандарты

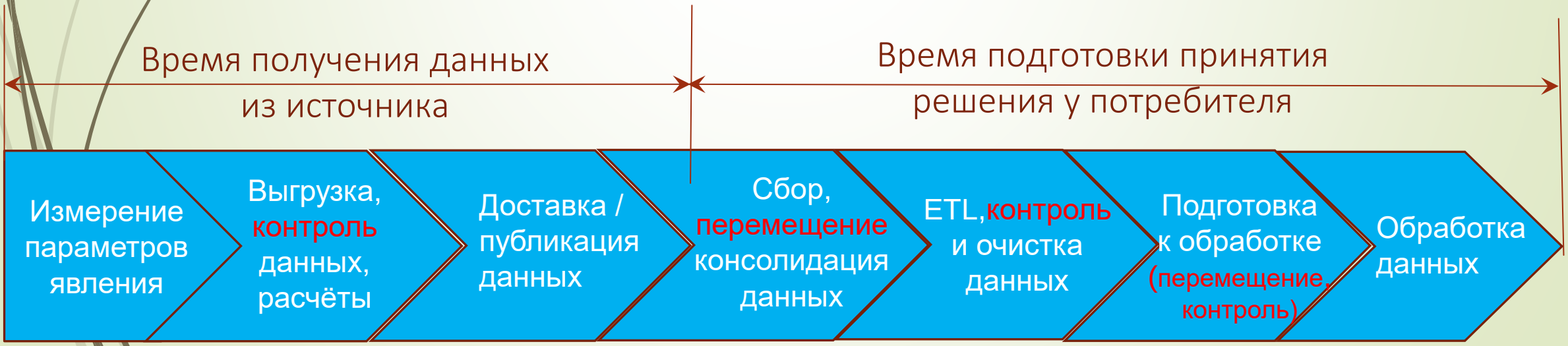


Профиль обеспечения своевременности данных

Наличие нескольких точек контроля качества данных – реальность.

Важно: минимизировать объём контроля путём выделения критических данных, исключить дублирование проверок при предоставлении протокола, распараллеливать контрольные процедуры.

Для сокращения / устранения перемещения данных использовать технологии: витрина данных источника, режим pull вместо push, логическое хранилище данных LDW, гибридная транзакционно-аналитическая обработка данных HTAP.



Характеристики релевантности данных

Доступные наборы и элементы данных соответствуют смыслу решаемой задачи?

Известны бизнес-правила?

Классификации соответствуют?

Соответствие смыслу задачи

Удовлетворяет требованиям задачи детальность доступных данных?

Достаточна точность данных для решения задачи?

Подходят единицы измерения?

Детальность данных

Существование необходимых данных

Имеется достаточный охват субъектов, объектов и явлений у доступных данных?

Имеются необходимые порции данных за требуемые периоды времени?

Применимость данных

Определены ли критические данные?

Имеются результаты профилирования и контроля данных?

Достаточен уровень качества данных?

12 Бизнес-глоссарий для оценки релевантности

Метрики расширенного качества данных

Доступность

Общее число используемых информационных ресурсов

Доля ресурсов
- учтённых в каталоге;
- интегрированных;
- доступных бизнес-пользователям;
- имеющих семантические слои пользователя;
- имеющих модели данных или описания форматов

Использование
- каталога данных,
- хранилища (витрин, озера) данных

Выполнение запросов доступа к данным

Своевременность

Нарушения SLA (сроков и периодичности):
- получения данных,
- подготовки данных (сбора, перемещения, консолидации, интеграции данных, контроля качества),
- обработки данных

Нарушение SLA процедур и бизнес-процессов

Продолжительность получения, подготовки, обработки данных

Профили подготовки принятия решений

Релевантность

Количество статей бизнес-глоссария

Число элементов данных
- критических,
- классифицированных

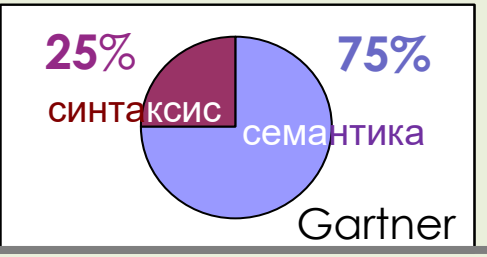
Доля критических элементов данных, увязанных с процессами

Доля активов данных с полным и ясным описанием

Использование бизнес-глоссария данных

Аналитика поиска в каталоге данных

Характеристики и показатели качества данных

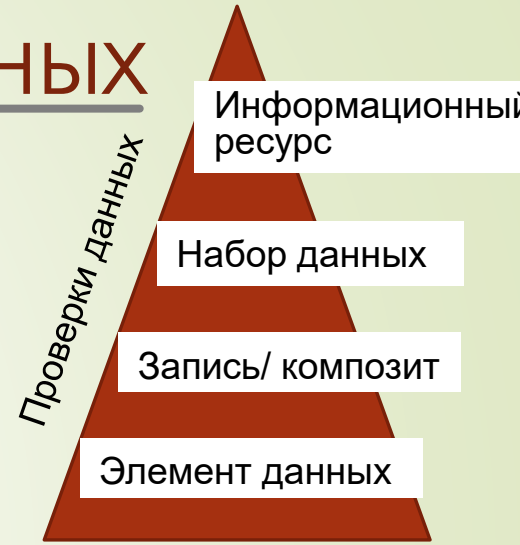


ДОСТОВЕРНОСТЬ – соответствие содержания и структуры данных реалиям

- **Полнота** – наличие непустых значений для элементов данных (композита), всех необходимых (или достаточных) записей или наборов данных
- **Допустимость** – соответствие отдельных элементов данных (или групп элементов данных в записи) *области допустимых значений*
- **Целостность структуры** – наличие связей наборов данных
- **Согласованность** – соответствие данных *бизнес-правилам* на уровне элементов данных, записей (композитов), наборов данных, информационных ресурсов

- тип данных
- разрядность и точность числа
- длина строки/ кода
- формат
- диапазон/ перечень допустимых значений
– контрольный разряд

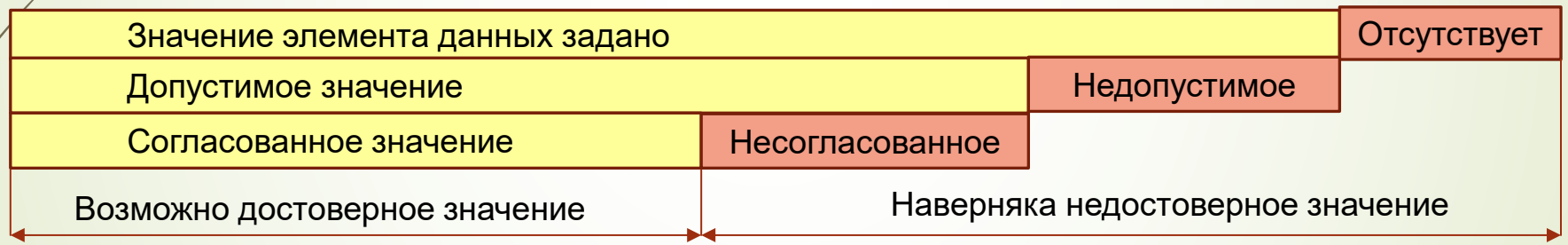
Проверки и метрики качества данных



Проверки данных

Для каждого критичного/ важного уровня данных разрабатывают подходящие проверки по полноте, целостности структуры, допустимости и согласованности.
Важно чётко формулировать суть проверки и относить к определённому показателю качества и уровню данных

Зависимость проверок

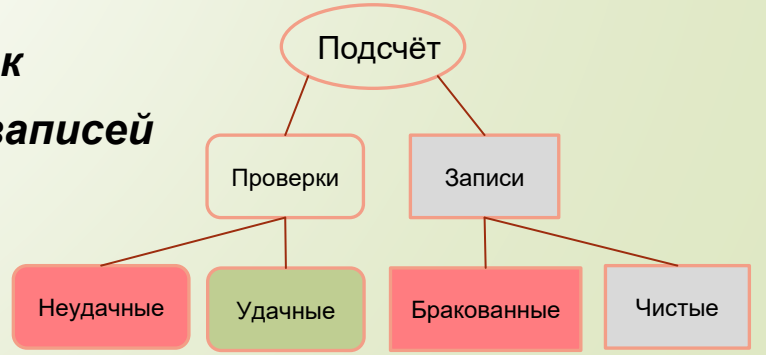


Метрики показателей качества

Доля неудачных проверок = ***Число ошибок*** / ***Общее кол-во проверок***

Доля бракованных записей = ***Число бракованных записей*** / ***Число записей***

Доля бракованных записей ≠ ***Доля неудачных проверок***



Начальные шаги анализа качества данных

60% - 80% подготовки принятия решений

- **Формирование требований к качеству данных**

  Изучение описаний критических и важных данных.
  Выявление схем, просмотр содержания данных.
  Выбор элементов и наборов данных для проверки

- **При недостатке сведений – профилирование данных**

  - Анализ частоты значений
  - Обнаружение шаблонов
  - Анализ области значений
  - Анализ структуры
  - Анализ избыточности

- **Выбор показателей, разработка тестов и метрик**

  Показатели выбираются в зависимости от формата данных.
  Определение детальных и интегральных метрик

- **Разработка и внедрение процедур контроля и метрик**
- **Контроль и анализ качества данных,** принятие решения об очистке
- **Очистка данных** и, возможно, повторный контроль

  - Исключение данных
  - Заглушки данных
  - Формальные правила
  - Регулярные выражения
  - Использование словарей
  - Таблицы замены
  - Стандартизация данных
  - Анализ сходства строк
  - Приведение к бизнес-правилам
  - Устранение дублей

- **Информирование о качестве данных**

  Публикация отчёта о качестве данных
  Или доступ к информации о качестве данных

Средства управления качеством данных (Data Quality Tools)

- **Отдельные инструменты управления качеством данных (корпоративные и open source)**
  - Анализаторы качества данных (профилирование)
  - Системы сбора данных
  - Средства очистки данных
  - Средства мониторинга качества данных
  - Информационные панели (dashboard)
  - Средства извлечения, преобразования и загрузки данных (ETL)
  - Запись GUI-манипуляций и скрипты
  - Языки программирования и библиотеки
- **Средства в рамках платформ интеграции данных (Data Warehousing)**
- **Средства в рамках платформ управления мастер-данными (MDM)**

Пример: Оценка качества Витрины данных торгового репозитария и депозитария НРД

- **Динамика доли ошибок витрины данных по показателям качества** М (ИР, показатель, год)

Год	Методологическая определенность	Согласованность	Достоверность	Полнота	Целостность	Общая доля ошибок
2015	293,2	19,4	18,3	1,5	0,9	15,4
2016	14,7	35,7	3,5	6,8	5,3	9,5
2017	0,4	12,5	4,3	4,6	4,5	5,1

промилли

- **Сравнение двух интегральных метрик наборов данных и витрины** М (ИР, НД)

Тип ПФИ	Число ошибок	Число проверок	Число записей	Число бракованных записей		Доля бракованных записей		Доля ошибок
				НГ	ВГ	НГ	ВГ	
Форвард на фондовые активы	7 397	43 266	7 211	7 211	7 211	100,00%	100,00%	17,10%
РЕПО	416 271	34 799 466	2 676 882	174 930	416 271	6,53%	15,55%	1,20%
Товарный форвард	1 246	137 520	22 920	929	1 246	4,05%	5,44%	0,91%
Валютный спот или форвард	90 602	18 739 568	2 342 446	86 351	90 602	3,69%	3,87%	0,48%
Валютный (конверсионный) своп	10 857	9 499 980	791 665	5 486	10 857	0,69%	1,37%	0,11%
ИТОГО	**526 373**	**63 219 800**	**5 841 124**	**214 829**	**526 187**	**4,71%**	**9,01%**	**0,83%**

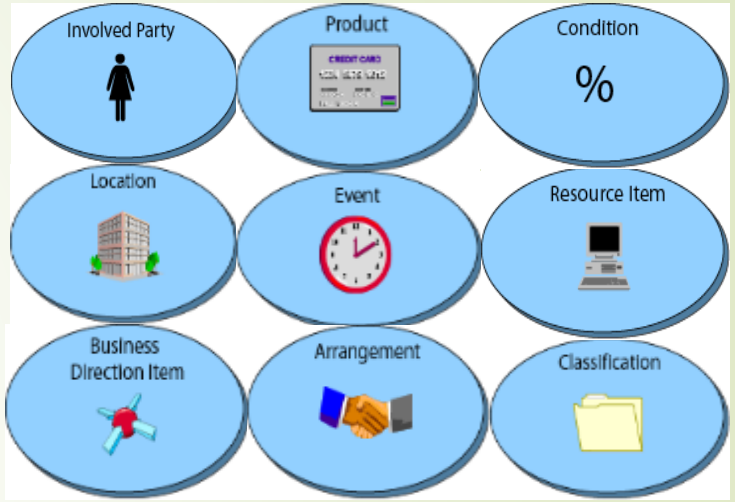
М (ИР)

НГ (***Доля бракованных записей***) = МАКС {***Кол-во ошибок по i-ой проверке***}/ ***Кол-во записей***

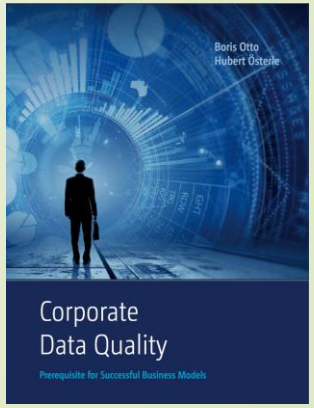
ВГ (***Доля бракованных записей***) = МИН (***Кол-во ошибок***; ***Кол-во записей***)/ ***Кол-во записей***

Управление справочными и реестровыми данными (Reference & Master Data Management)

- **Реестры и справочники** – критические и важные информационные ресурсы
- **Реестры – данные о ключевых бизнес-сущностях**
  - Организации, ИП и физлица
  - Местонахождение и контакты
  - Продукты и услуги
  - Договоры
  - События, операции и т.п.



- **Обеспечение качества мастер-данных**
  - Унификация идентификации субъектов
  - Формирование «золотых» записей (история изменений, кластеризация, оценка достоверности, приоритеты заполнения)
  - Стандартизация (парсинг, правила, шаблоны, словари, matching, определения вида субъекта)
- **Классификаторы – аналитическая ценность данных**
  - Международные, российские, отраслевые, ведомственные и локальные
  - Устойчивость и полнота схем классификации
  - Методы классификации и кодирования



Дедубликация и обогащение реестра ЛИЦА

- Источники (в порядке важности)
  - ЕГРЮЛ/ ЕГРИП (ФНС)
  - Инфоресурс СПАРК-ШЛЮЗ
  - КГРКО (Банк России)
  - РУФР (Банк России)
  - ЕРСМП (ФНС)
  - БИК и SWIFT (Банк России)
  - Статрегистр (Росстат)
- Правила кластеризации данных источников

Поля кластеризации	Пояснение
ИНН, ОГРН/ОГРНИП + дата ОГРН/ОГРНИП	Сильное правило по двум основным идентификаторам и дате
ИНН, ОГРН/ОГРНИП	Ослабленное правило по двум основным идентификаторам без даты
ИНН + дата ИНН или ОГРН/ОГРНИП + дата ОГРН/ОГРНИП	По одному из основных идентификаторов и его дате
ИНН, КПП или ОГРН/ОГРНИП, КПП	По одному из основных и дополнительному идентификатору
ИНН или ОГРН/ОГРНИП	По одному из основных идентификаторов

Оценка качества ЕГРЮЛ

Интегральные метрики

Характеристика	Доля ошибок	НГ брака	ВГ брака
Согласованность	4,83%	28,55%	28,96%
Полнота	3,24%	19,38%	45,35%
Целостность	1,32%	1,32%	1,32%
Допустимость	0,04%	0,12%	0,12%
Общий итог	**3,16%**	**28,55%**	

Доля ошибок =
Число ошибок / Число проверок

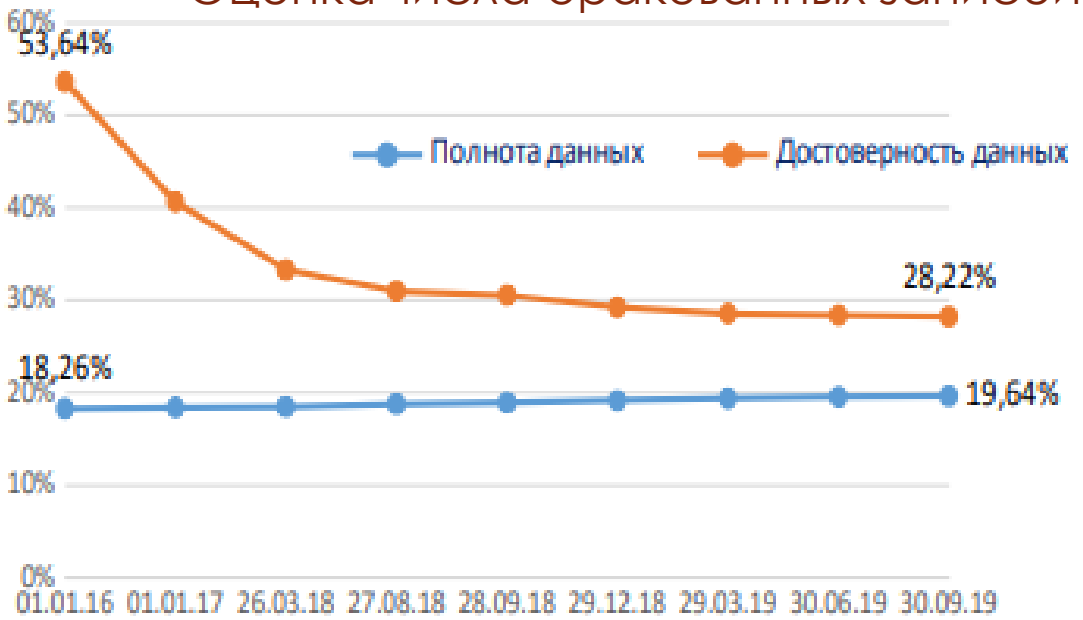
Проверки и детальная метрика

Характеристика и проверки	Число ошибок	Число проверок	Доля ошибок
Согласованность	**1 056 826**	**21 898 128**	**4,83%**
Сумма долей в УК в точности равна 100%	1 042 056	3 649 688	28,55%
Дублирующиеся лицензии отсутствуют	10 343	3 649 688	0,28%
Основной код ОКВЭД ровно один	4 293	3 649 688	0,12%
Совпадает КПП в записи об учете в налоговом органе	102	3 649 688	0,0028%
Совпадает ИНН в записи об учете в налоговом органе	16	3 649 688	0,0005%
Ровно одна запись по сочетанию ИНН и ОГРН	16	3 649 688	0,0005%

Сравнение метрик качества данных ЕГРЮЛ

Доли ошибок при проверке данных

Оценка числа бракованных записей



Оценка качества Реестра производителей «ювелирки»

Статистика (на 01.072018)

Реестр	Брак	Записей	Процент	Интервал
ЮЛ	198	903	21,9%	4,6
ИП	593	3214	18,5%	5,4
Производители	**791**	**4117**	**19,2%**	**5,2**

Наибольшее число брака из-за несогласованности с Общим реестром и нарушением уникальности

Примеры перегруженных элементов данных (от 2 до 6):

Уч.номер - **0080007782** ИНН **5403172040**
№ 0140008560 (**ЮЛ7801401687** с 18.01.2018 от 02.02.2018 №30-14-02-03/117) ИНН **7806104390**

Курьёзы наименования:

- «…общество с ограниченной возможностью»
- «…общество с ограниченной общественностью»
- «…общество с ограниченной отнесенностью»
- «…общество с ограниченной отвлечённостью»

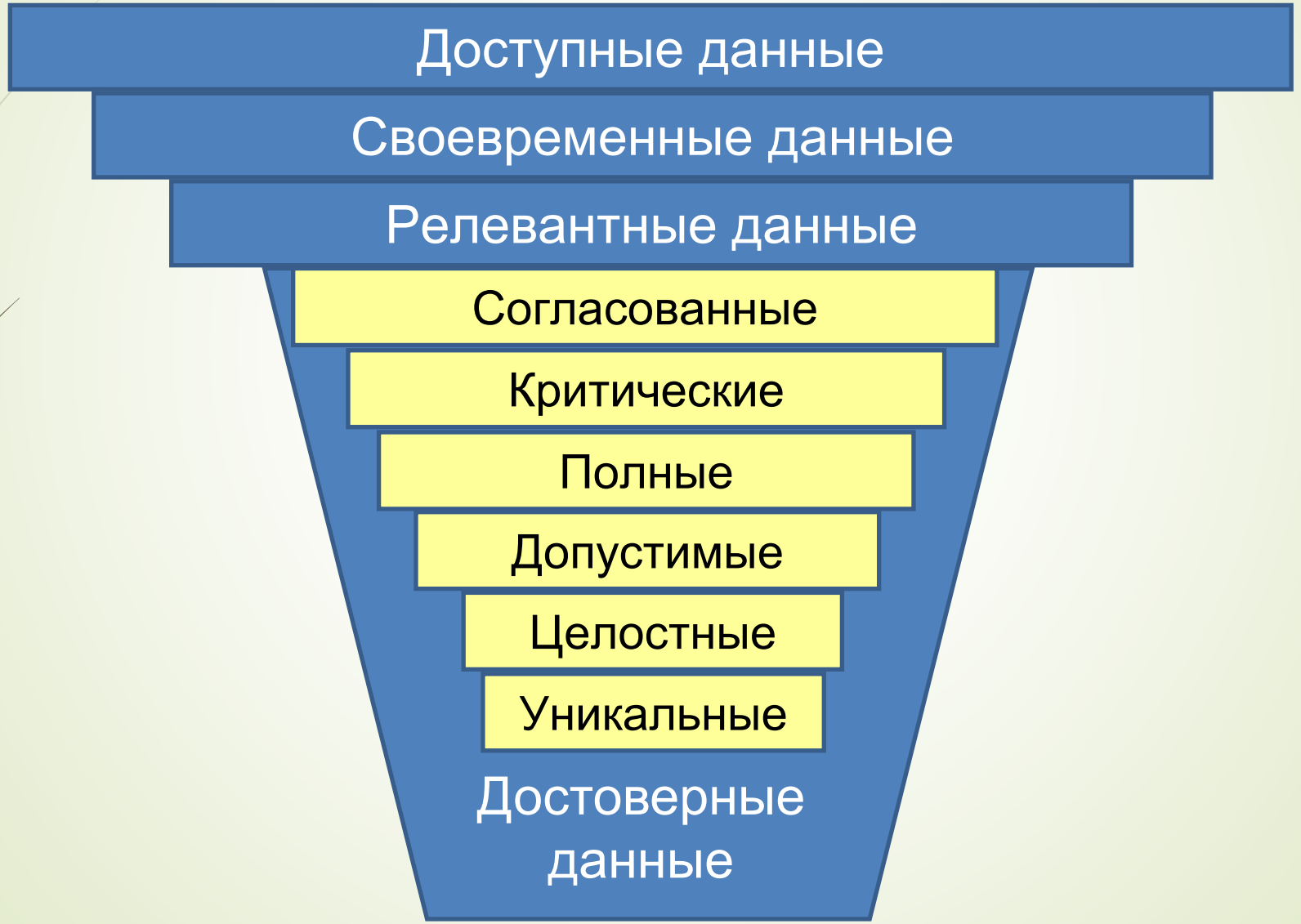
Конвейер обеспечения достоверности данных

- Сравнение с уровнями качества
- Извещение о нарушении DQA

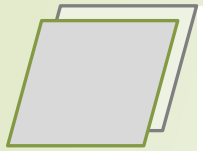
- Фиксация ошибок данных
- Подсчёт ошибок данных
- Расчёт метрик достоверности

- Публикация отчёта о качестве
- Доступ к метаданным о качестве
- Связывание с каталогом данных

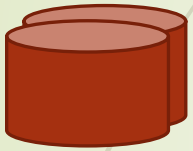
Воронка расширенного качества данных



Расширение спектра информационных активов



«Сырые» данные
Собираемая отчётность и микроданные, события, сообщения, измерения с датчиков, данные API/web-сервиса,



Информация
Фактографические данные, реестры, классификаторы/ справочники, результаты обработки, документы, контент, архивы

Метаданные
Описания информационных активов, модели данных, технические метаданные

Знания
Семантика предметных областей, описательные и прогнозные модели, имитация, рекомендации, отношения/связи, извлечение фактов и выявление тональности текста

Артефакты искусственного интеллекта
Предписывающие модели, распознавание/ генерация письменной и устной речи, распознавание образов, поиск и запросы на естественном язык

Управление метаданными (Metadata Management)

- **Каталог данных – реестр информационных ресурсов**

  Назначение данных/ информации, уровень доступа, источники, владельцы/ кураторы, поставщики, потребители, ресурс/ приложение, ссылка на прикладную модель данных, качество данных

- **Корпоративная модель данных – репозиторий моделей**

  Верхнеуровневая модель данных, концептуальные и логические модели предметных областей, логические и физические прикладные модели, связи с бизнес-глоссарием. Каноническая модель данных

- **Бизнес-глоссарий данных – семантика данных**

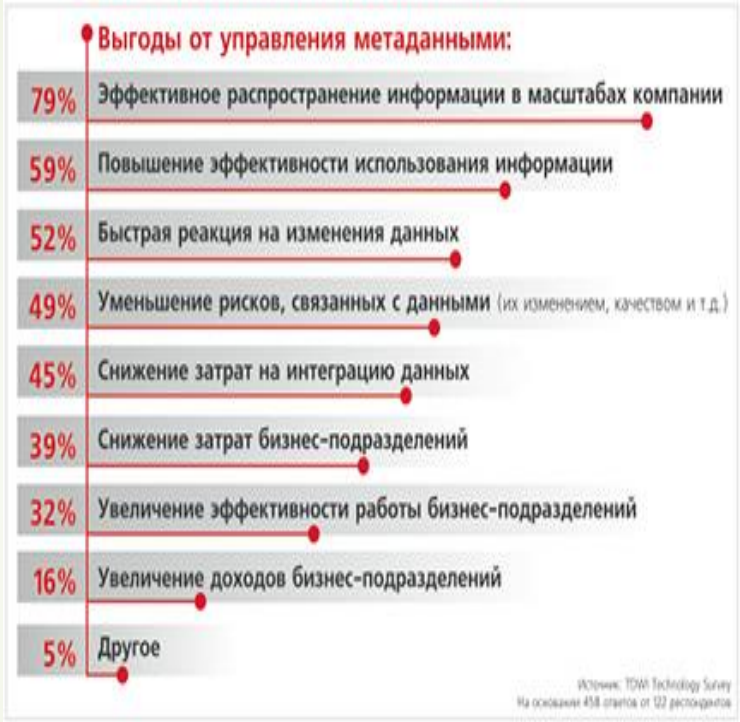
  Предметные области данных, сущности, связи, атрибуты, области допустимых значений, бизнес-правила, ссылки на нормативную базу и тезаурус

- **Каталог аналитических моделей и артефактов ИИ**

  Назначение модели, структура модели, метод, алгоритм, приложение/ инструмент, обучающий и тестовый наборы, параметры, вход/ выход, качество модели

- **Общий репозиторий метаданных**

  Содержит или связывает все метаданные, включая дополнительные



Моделирование и проектирование данных (Data Modeling & Design)

- **Модель данных** – описание структуры и содержания данных для представления реального объекта, процесса или концепции
- **Понятность и повторная используемость данных**
- **Операции**
  - Ведение моделей
  - Версионирование
  - Связывание
  - Навигация и поиск
  - Импорт/ экспорт
- **Инструмент**
  - SAP Power Designer

Владельцы/ кураторы данных
Модельеры, архитекторы
Аналитики, спецы по качеству данных
Проектировщики, программисты
Администраторы БД

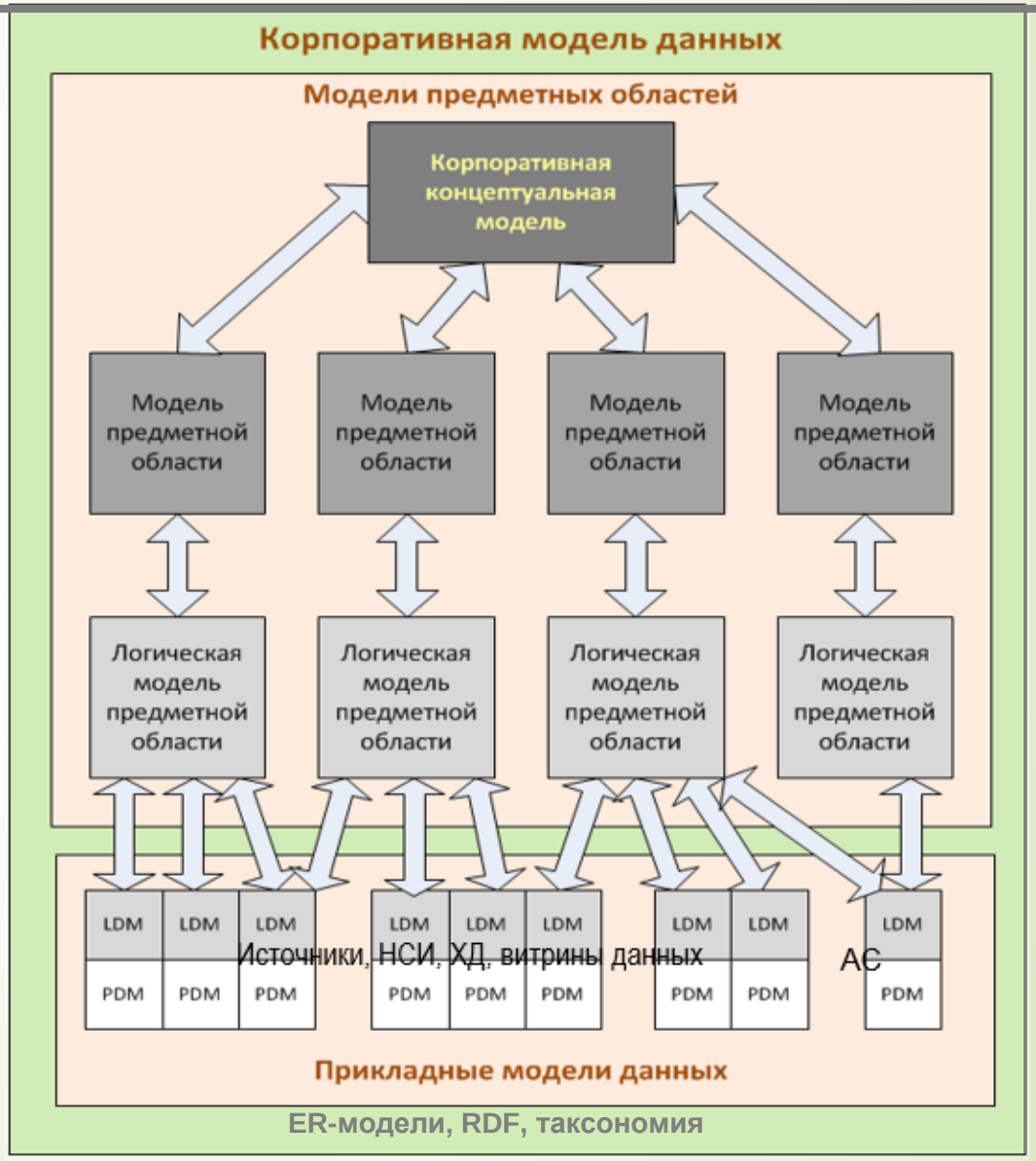
Концептуальный обзор всех предметных областей корпорации

Концептуальное представление сущностей и связей для каждой области

Логическое представление для каждой предметной области

Каноническая модель данных: форматы сообщений и web-сервисов (XML / JSON)

Подробные логические и физические модели данных, специфичные для реализации приложений или проектов



Хватит использовать «грязные данные»!

- **Помните, что качество данных зависит от их применения**
- **Определяйте критичность и важность наборов и элементов данных**
- **Измеряйте качество используемых данных, для чего:**
  - Правильно выбирайте показатели качества данных
  - Корректно формулируйте проверки и распределяйте по показателям
  - Рассчитывайте детальные и интегральные метрики как долю бракованных записей
  - Наблюдайте динамику качества данных
  - Раскрывайте сведения о качестве данных для потенциальных пользователей
- **Планируйте необходимую очистку данных**
- **Уделяйте основное внимание качеству реестров и классификаторов**
- **Метаданные важны для понимания и организации контроля**
- **Модели данных, бизнес-глоссарий и бизнес-правила – семантическая основа качества данных**
- **Учитывайте информационные ресурсы, аналитические модели и артефакты ИИ**
- **Повышайте осведомлённость и развивайте компетенции по качеству данных**