

Малые данные + генеративный ИИ

01 ноября 2023

Гуреев Дмитрий

2 слова о себе

Гуреев Дмитрий

(МВА, выпускник CDTO-3 Сколково, соавтор курса «ИИ для бизнеса», приглашенный эксперт по работе с данными в МШУ Сколково)

CDTO Биовитрум

- 18 лет в области лабораторной медицины (со стороны поставщиков)
- Карьерный путь от сервисного инженера до генерального директора
- Обширный операционный опыт, в т.ч. антикризисный: от продаж до R&D
- Цифра всегда помогала делать карьеру

Исследование:

- Управление малыми данными как сквозной навык
- Роль локальных LLM в построении бизнес – процессов нового типа





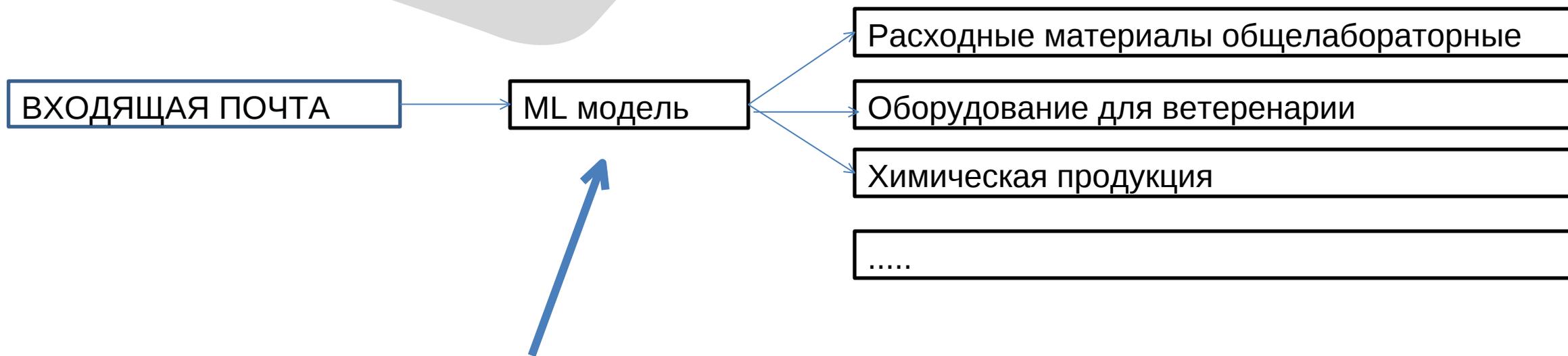
Малые данные есть у всех.
Большие у единиц.

Малые данные (МД) - оперативные и справочные бизнес-данные, хранящиеся в информационных системах.

Что бизнес хочет от малых данных?

- Дашборды в один клик из всех систем и интеграция всего со всем.
(интероперабельность)
- Чтобы не было ошибок в данных и в них можно верить.
(валидность и консистентность)
- Доступ к данным был открытым внутри и закрытым извне.
(ИБ)
- На их основе создавать микросервисы.
(различные задачи автоматизации/цифровизации)

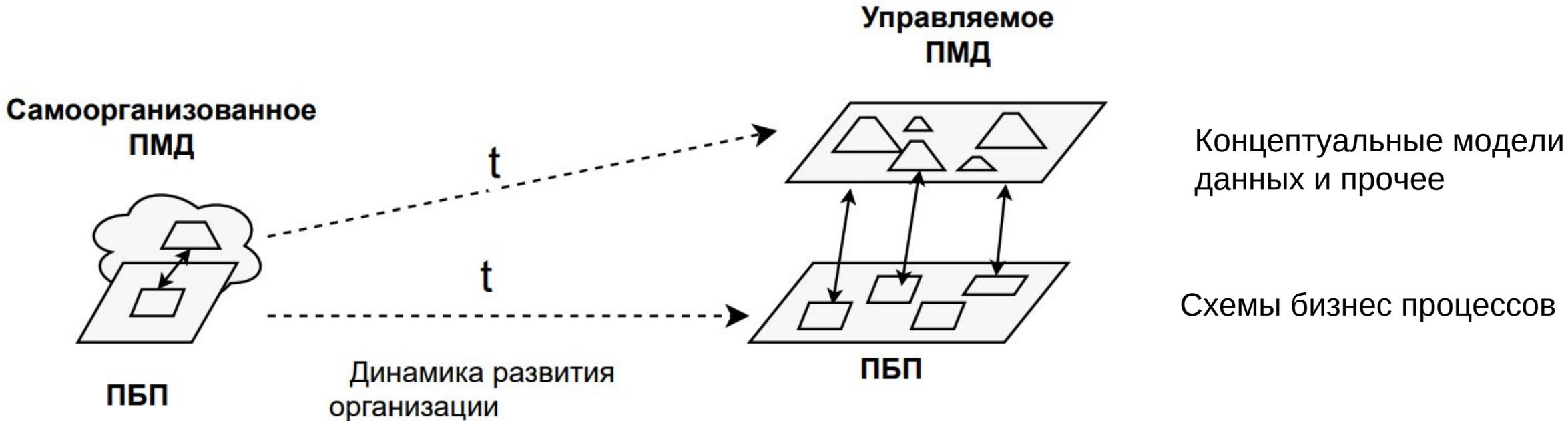
Классификация входящих запросов клиентов по товарным группам



Классификационная модель натренированная на справочнике товаров из 1С (малые данные)



Автоматизируя хаос, получим
автоматизированный хаос



Определения

- **Малые данные (МД)** - оперативные и справочные бизнес-данные, хранящиеся в информационных системах.
- **Пространство малых данных (ПМД)** - совокупность всех бизнес-данных, которые создаются, преобразовываются и хранятся в контролируемых организацией информационных системах.
- **Пространство бизнес-процессов (ПБП)** - совокупность функций (закупка, маркетинг, продажа, производство...), которые делает компания в рамках своего основного бизнес-процесса.

**Большие компании создают отделы
Data Office,
средние компании – в лучшем случае
чистят справочник контактов.**

Дорого, долго и сложно.
А можно как-то иначе?



**LLM (LARGE LANGUAGE MODELS)
CHATGPT, YAGPT, GIGACHAT, LLAMA2,
MISTRAL.....**

**ЭТИ МОДЕЛИ ЭВОЛЮЦИОННО УЛУЧШИЛИ ПРИНЦИП КОММУНИКАЦИИ
ЧЕЛОВЕК – МАШИНА, ДОВЕДЯ ЕГО ДО СТИЛЯ
ЧЕЛОВЕК - ЧЕЛОВЕК**

Оценка рынка в базовых единицах (кг,литры) через LLM

Дано (из базы тендеров):

текстовая строка с названием медицинского товара и возможно с указанием размерности упаковки

Необходимо:

дообогатить строку столбцом с указанием сколько базовых единиц товара

Парафин для гистологической заливки Гистомикс , 5кг - Сколько кг?

Формалин 15x25мл в кор. 5 кор/уп - Сколько литров?

Ограничение:

- типов единиц и типов товаров неограниченно
- все записи различны, структуры записей нет и не будет
- есть товары (и их большинство), по котором нет экспертизы

LLM справились одним промптом

Ты технический писатель. Определи сколько литров в ... (промпт сократил)

Формалин 15x25мл в кор. 5 кор/уп.

X GPT 3.5 : 3.75 л.

X YandexGPT: 5

V GPT4: 1.875

Профилирование данных в каталоге

- На входе примеры записей, на выходе тип данных (метаданные):
 - Адрес
 - Телефон
 -

Text to SQL (лучший дашборд, которого нет)

The screenshot shows a Telegram chat interface. At the top, a message from the bot `@rusmedfile_assitance_bot` asks for the top 10 suppliers for a specific customer. Below this, a user named Dmitry Gureev sends a SQL query. The bot's response lists the top 10 suppliers with their names and sales for 2022 and 2023.

`@rusmedfile_assitance_bot` покажи топ 10 поставщиков продукции покупателю с ИНН 7720024671, которые за период с января по июль 2023, продали ему больше по сумме товаров, чем за период с января по 2022. Укажи суммы их продаж за 2022 и 2023 года и названия. 18:44 ✓✓

Ассистент Русмедфайл Reply
Dmitry Gureev
`@rusmedfile_assitance_bot` п...
Пошел искать в базе. 18:44

Dmitry Gureev
`@rusmedfile_assitance_bot` покажи топ 10 поставщиков проду...
supplier_name -- sum_product_total_sum_2022 --
sum_product_total_sum_2023

ООО "ХИМИЯ И ЖИЗНЬ" ; 12124614.06 ; 21426600

ООО "СКАЙДЖИН" ; 11409330 ; 17619736.8

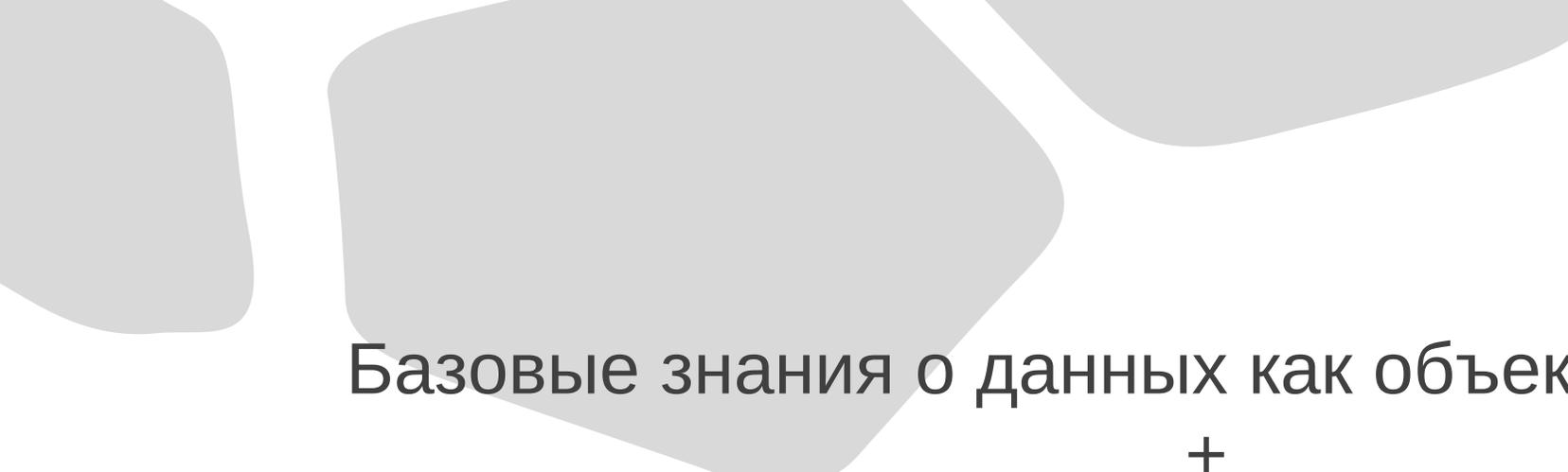
ИП Бганба Леонид Беспанович ; 0 ; 10538898.36

ООО "МОЛБИОЛ" ; 0 ; 5877617.72

ЗАО "БИОХИММАК" ; 0 ; 3232500.7

ООО "ДИАВЕНДОР" ; 0 ; 2828400

ООО "МП-ГАРАНТ" ; 0 ; 2240056



Базовые знания о данных как объекте управления

+

LLM модели (+ Low Code)

=

Возможность управлять данными на эволюционно
ином уровне (быстрее, дешевле, проще)

ОСНОВНЫЕ ВЫЗОВЫ

- Перенос LLM моделей в закрытый контур, так как соблазн использовать OpenAI – велик
- Удобный и быстрый Fine-Tune моделей
- Углубляющееся цифровое неравенство (знаний)

Время вопросов и ответов

[Гуреев Дмитрий Владимирович](#)
dvgureev@gmail.com

+ 7 (925) 741 94 53

