КОНГРЕСС ОИРОМ

ТЕЗАУРУС СИНТЕТИЧЕСКИХ ДАННЫХ

Илья Буханский
эксперт Научно-методического департамента, АЦ ВЦИОМ

Максим Акульшин
генеральный директор SURVEYSTUDIO

Когнитивные искажения: идол рынка

Нечёткость соответствия между определениями слов и явлениями реальности, этими словами представляемыми

Порождает терминологическую неоднозначность, недопонимание

Делает из Дискурса Базар

Задача – перевести Базар в Дискурс

Метод – компиляция источников
+ опыт

Синтетические данные – искусственно созданные данные, имитирующие данные, которые обычно формируются естественно, иначе говоря, собираются от реальных источников

!Качество синтетических данных – внешняя валидность и внутренняя консистентность

1. Что имитирует система

Имитационная система (= синтетик) = система, имитирующая реакцию реального источника на состояния окружающей среды

Система может

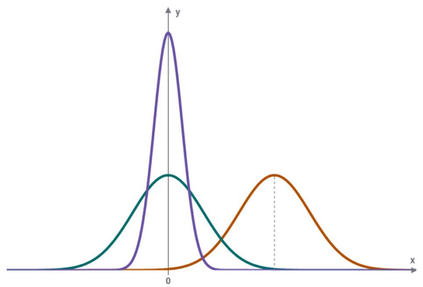
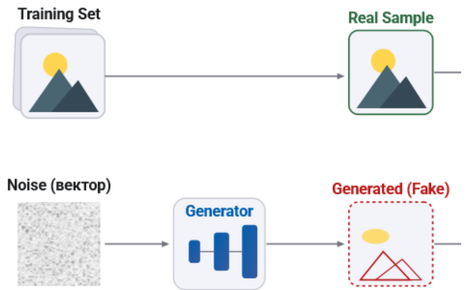
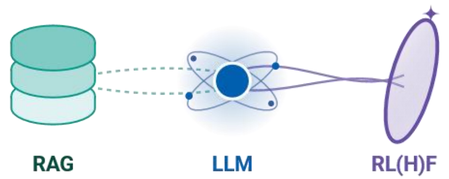
ИЛИ

1. Воспроизводить распределения (= групповой уровень)

ИЛИ

2. Реконструировать рассуждающего субъекта
   = синтетический респондент (= индивидуальный уровень)
   = мозг **без тела**

2. Что получаем в результате синтеза

Реальные данные + синтетические столбцы	Реальные данные + синтетические строки	Полностью синтетические данные
		

2. На каких реальных данных учится имитационная система

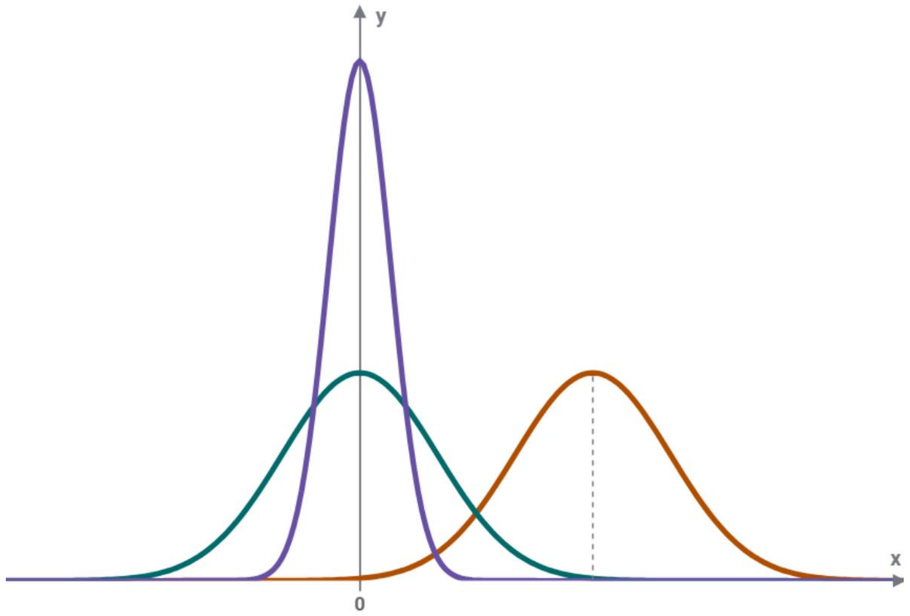
	Знания о реальном источнике, его реакциях	Знания об окружающей среде
Наборы данных (первичные данные)		
Знания, собранные аналитиками (вторичные данные)		
Гибрид		

!Переход из знания об окружающей среде в знание о реакции источника: все ли необходимые данные учтены?
!Консистентность нарративов об источнике и об окружающей среде

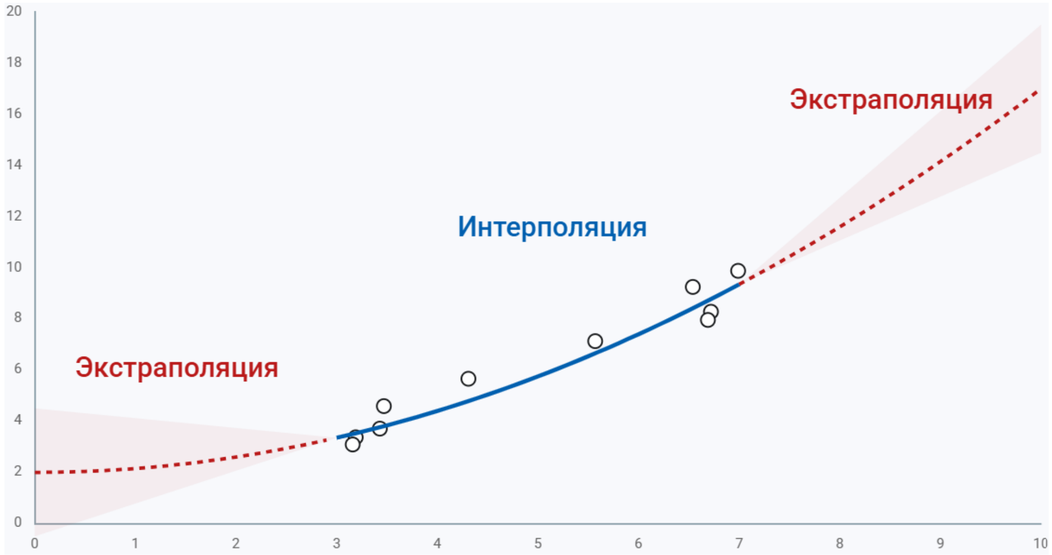
3. Какой алгоритм используется для извлечения закономерностей из данных

1) Воспроизводство агрегатов (статистики)

На основе распределений



На основе корреляций



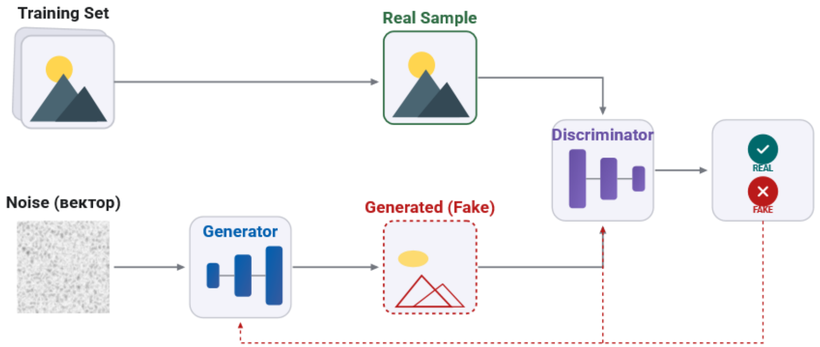
3. Какой алгоритм используется для извлечения закономерностей из данных

2) Генерация строк на основе совместного распределения переменных

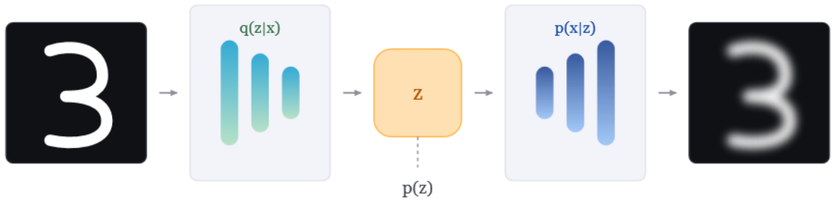
ML

+

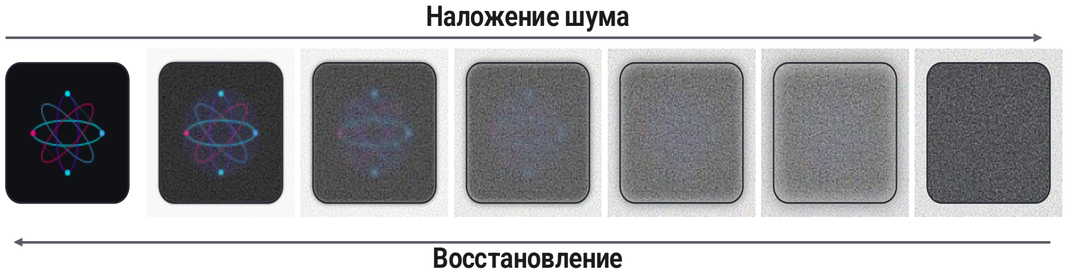
Состязательные сети (GAN)



Автоэнкодеры (VAE)

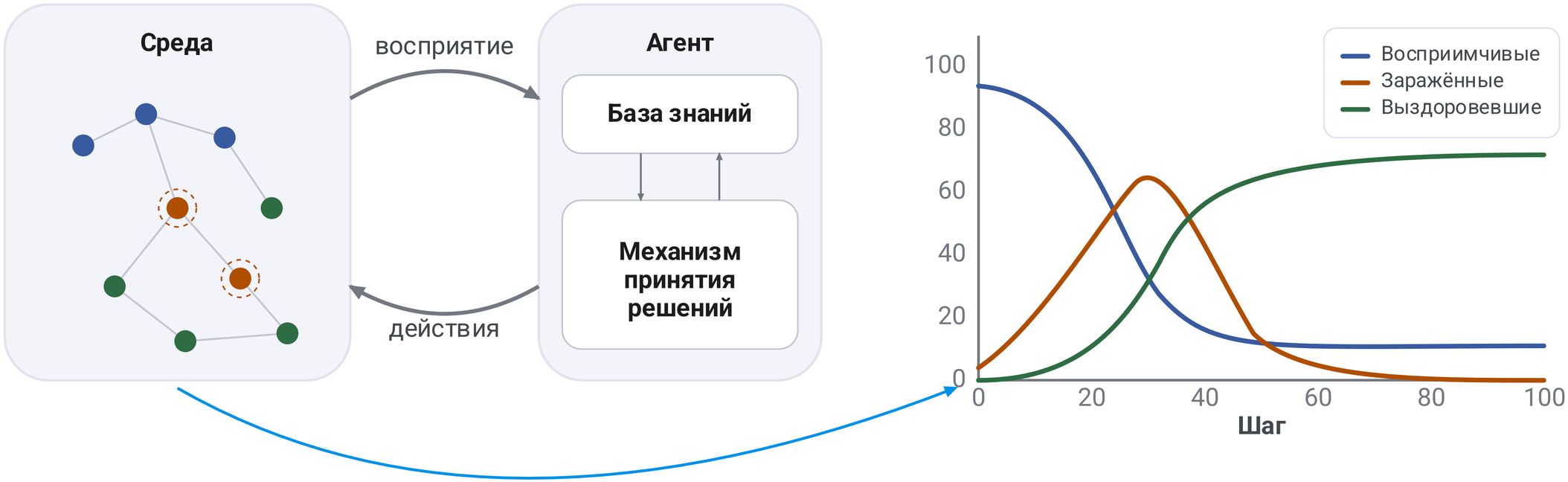


Диффузионные модели



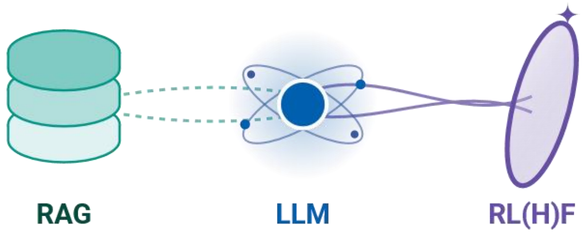
3. Какой алгоритм используется для извлечения закономерностей из данных

3) Симуляция процесса (агентное моделирование)



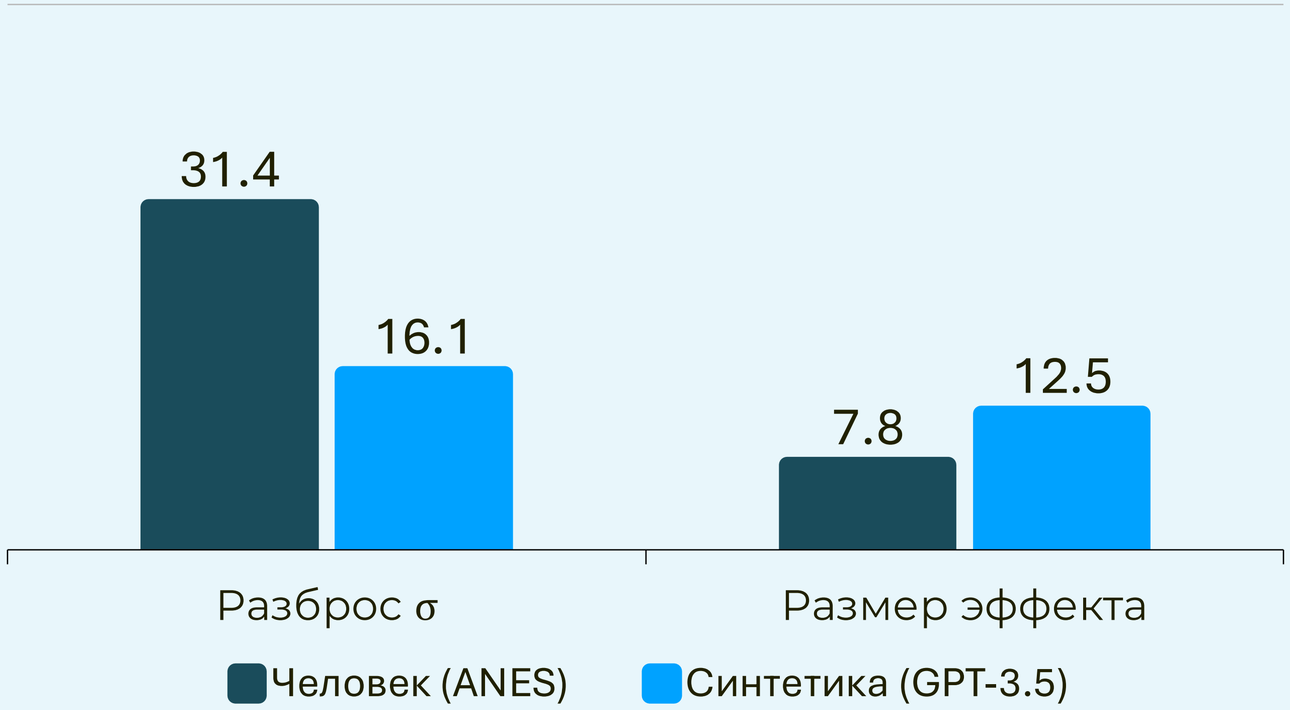
3. Какой алгоритм используется для извлечения закономерностей из данных

4) Реконструкция субъекта (синтетический респондент)
(+ его опрос для получения синтетических данных)



LLM	Строит «модель мира», выучивает процедуру размышления, принятия решений из данных в форме текста (через решение множества текстовых задач) Требует целостных, детализированных, новейших, многоуровневых/глубоких данных Обязательна калибровка на все группы, в том числе мало представленные
RAG	Контролируемая база знаний, из которой LLM при ответе достаёт релевантные вопросу данные (актуализация) Хранит данные в виде, в котором работает с текстами LLM => позволяет быстро искать (семантически) близкую информацию к той, что в работе у LLM
RL(H)F	То же самое, что у состязательных сетей, но обычно с человеком в качестве дискриминатора

Человек (ANES) vs синтетика LLM (GPT-3.5)



Стандартное отклонение, σ

31,4 → 16,1

↓ σ сжато на 49%; дисперсия σ^2 - до $\approx {}^1/_4$ от человеческой

Размер эффекта (поляризация)

7,8 → 12,5

↑ оценка завышена на 60% - ложная значимость

Почему ломается оценка погрешности: измеренный разброс – артефакт сэмплинга (temperature, top_p), а не свойство популяции. Формула $\sigma/\sqrt{n}$ даёт ложно узкий доверительный интервал, который при росте числа синтетических «респондентов» стягивается к смещённому среднему, а не к истинному (Chen et al., 2025).

Что с этим делать

Привязать обильную синтетику LLM к небольшой человеческой выборке и применить статистическую поправку

Человеческая выборка

маленькая, «золотой стандарт»

Синтетика LLM

обильная, но смещённая

Статистическая поправка

PPI • CDI • DSL

+ mixed-subject design

исправляет смещение модели по маленькой реальной выборке → несмещённая оценка

Валидный доверительный интервал

узкий И корректный

Ускоритель, не замена

доверительный интервал никогда не хуже, чем по одним людям

Нужен реальный якорь

валидной погрешности из чистой синтетики не существует

Меньше бывает лучше

оптимальное число k часто однозначное; больше синтетики не значит точнее

4. Вызовы, стоящие перед синтетическими данными

1. Вызовы, связанные с обучающим набором данных:
    - Устаревание обучающих данных = усвоенных моделями алгоритмов;
    - Коллапс знания (аппроксимация аппроксимации);
    - Смещения, присутствующие в реальных данных;
    - Риски безопасности данных;
2. Верификация (на прошлом и будущем, всё вне тренировочного массива);
3. Невозможность посчитать ошибку выборки.

Ясность ответа на вопрос
«Что имитирует система?»
- результат определение места предложенного продукта в описанных классификациях:

1. **Что получаем на выходе из модели (по степени «синтетичности»);**
2. **Какие данные модель получает для обучения и работы;**
3. **Какой (или какие) алгоритм лежит в основе извлечения закономерностей из подаваемых данных;**
4. **Как решаются вызовы, свойственные всей синтетике.**

ТЕЗАУРУС СИНТЕТИЧЕСКИХ ДАННЫХ

АЦ ВЦИОМ
Буханский И.
Оберемко О.