

Гарин Е.В. Необходимые и достаточные условия изобретающей машины

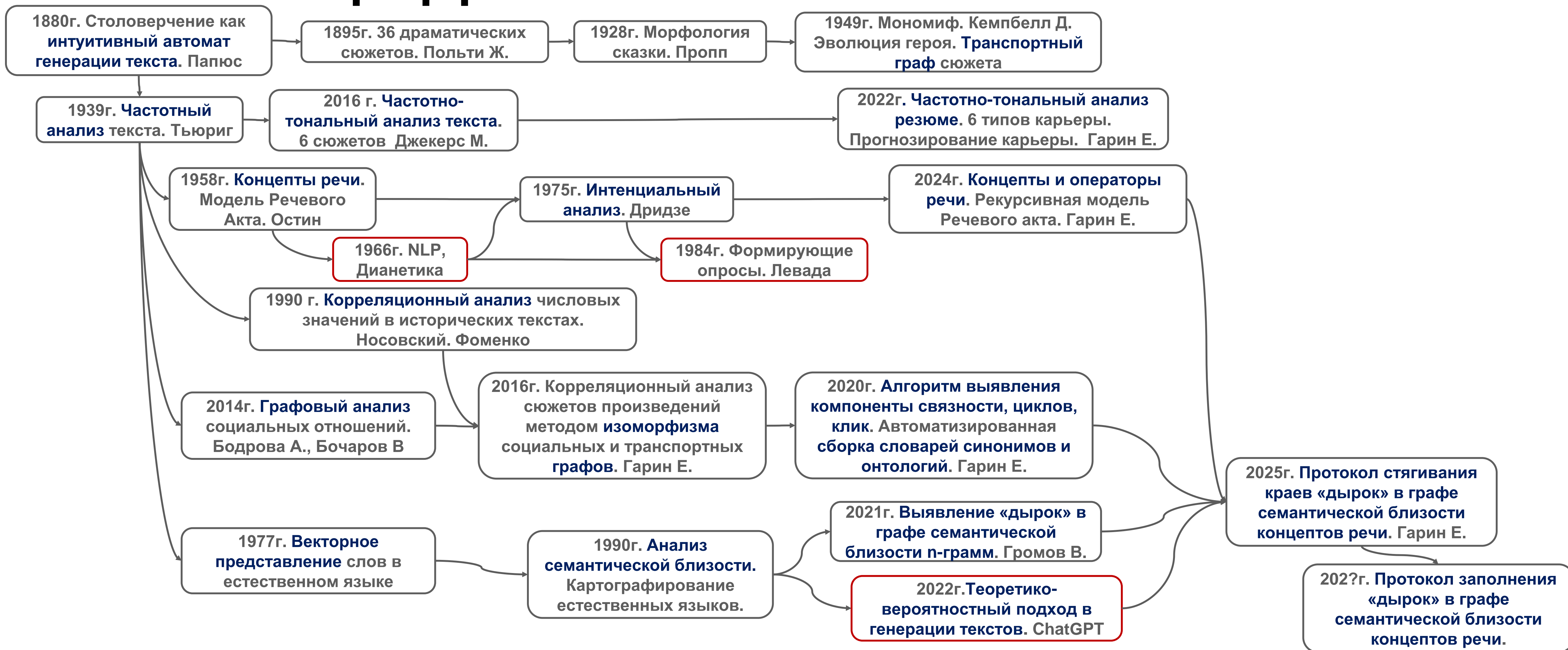
Ключевые слова: изобретающая машина, концепты и операторы речи, семантическая близость концептов речи, лакуны семантической близости.

Аннотация: В данной работе процесс научного открытия представлен в виде последовательности решения подзадач: 1) сборка корпуса текстов в научной стилистике; 2) представление корпуса текстов в виде графовой базы данных (тензорной БД) – графа семантической близости концептов речи; 3) выявление в базе данных лакун семантической близости; 4) заполнение лакун методами изоморфизма графов либо стягивание краев лакун путем комбинаторной замены концептов речи по критерию семантической близости. Каждая из четырех подзадач представляет собой конечный автомат. При этом совокупно последовательное решение всех четырех подзадач позволяет перейти к автогенерации новых знаний: семантических полей без лакун – то есть автоматически без участия человека находить пробелы в знаниях, парадоксы, противоречия и разрешать их.

Оптимальная архитектура изобретающей машины



Граф развития семантического анализа



Поисково-аналитическая платформа INVENTORUS



Технологии
OSINT*

Сбор

Обработка

Обогащение данных



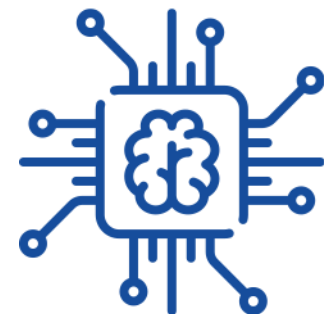
Структурированная
база данных
INVENTOSCI

Автономность

Безопасность

Аналитика

500 млн.ед НТИ



AI+COPILOT*

Анализ

Интерпретация

Предиктивная аналитика



патентов
154 млн



публикаций
256 млн



авторов
92 млн



организаций
12 млн



Поисково-аналитическая платформа "INVENTORUS" предназначена для поиска, анализа и интерпретации открытых данных о технологиях с использованием искусственного интеллекта, облегчает доступ к знаниям и поиску научно-технической информации со всего мира, позволяет анализировать и отслеживать технологические тренды, выявлять центры компетенций и исследовать технологические области при помощи искусственного интеллекта.

Входит в Единый реестр российских программ для электронных вычислительных машин и баз данных (Реестровая запись №21474 от 08.02.2024).

База данных "INVENTOSCI" имеет Свидетельство о государственной регистрации №2024620207.

Получена аккредитация IT компании в Министерстве цифрового развития, связи и массовых коммуникаций Российской Федерации (№ АО-20240301-16811465367-3 от 25,03.2024 г.)

RAPPEED – АНАЛИТИЧЕСКАЯ ПЛАТФОРМА ДЛЯ МАКСИМАЛЬНО ЛЁГКОЙ И БЫСТРОЙ РАБОТЫ С ДАННЫМИ ЛЮБОГО РАЗМЕРА, ДЕТАЛИЗАЦИИ И СТРУКТУРЫ

Ключевые преимущества:



Производительность – in-memory обработка расчетов на десятках миллиардов записей за считанные секунды



Масштабируемость – разворачивается в контейнерах с возможностью неограниченного горизонтального масштабирования



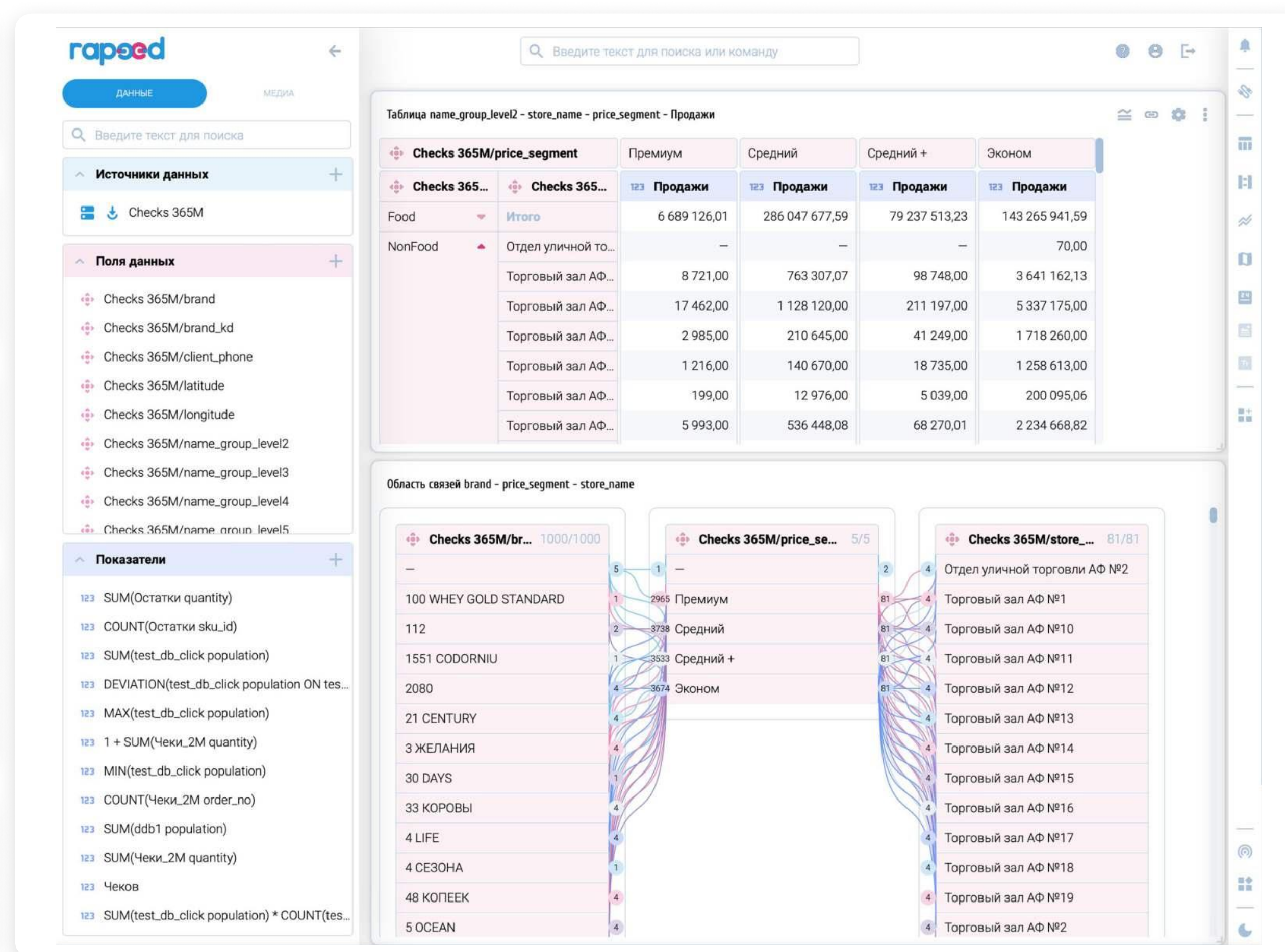
Динамичность - конфигурация данных в памяти автоматически подстраивается под запросы пользователей, без предварительных расчетов, JOIN-ов и агрегаций



Истинное самообслуживание – подготовка и изменение OLAP-кубов самостоятельно бизнес-пользователями. 100% аналитических запросов решаются в интерфейсе без написания SQL



Универсальность – единый аналитический слой данных для всех аналитических сценариев вашей организации с гарантированной консистентностью показателей



ДИНАМИЧЕСКАЯ РАСПРЕДЕЛЁННАЯ ТЕНЗОРНАЯ АРХИТЕКТУРА

■ Динамическая

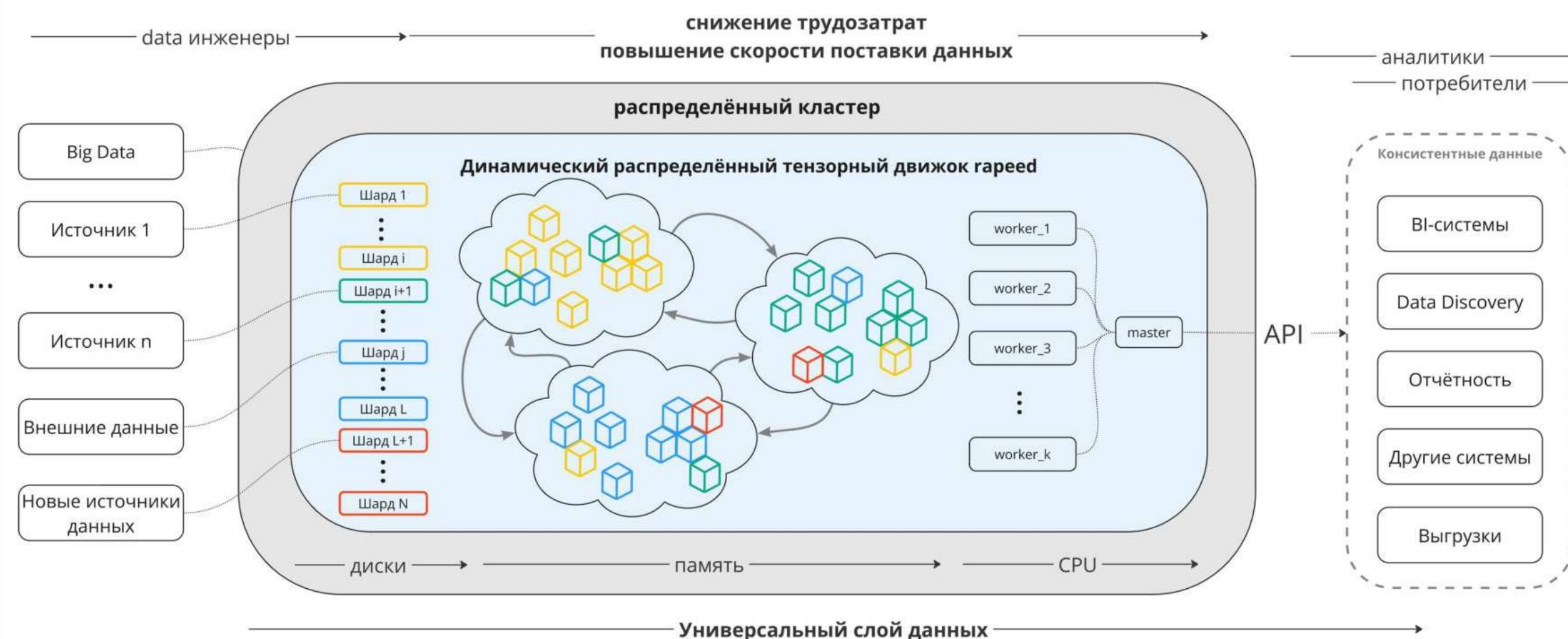
Нет никаких предварительных расчетов, JOIN-ов, агрегаций или других операций, видоизменяющих исходные данные; тензорный образ данных динамически формируется запросами пользователей

■ Распределённая

Движок разворачивается в контейнерах с возможностью неограниченного горизонтального масштабирования, при этом прирост производительности опережает прирост мощностей как \sqrt{n} (где n = число строк)

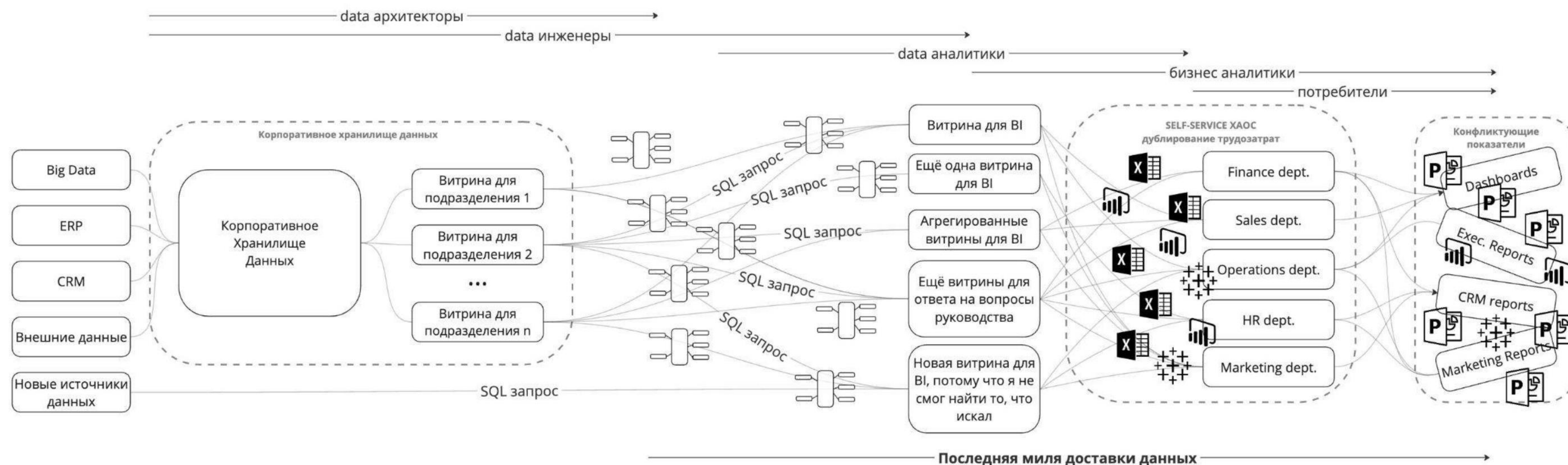
■ Тензорная

Данные из источников преобразуются в тензоры, вычисления обрабатываются тензорными операциями



КАКИЕ ОГРАНИЧЕНИЯ ТРАДИЦИОННЫХ OLAP-СИСТЕМ СНИМАЕТ RAPEED

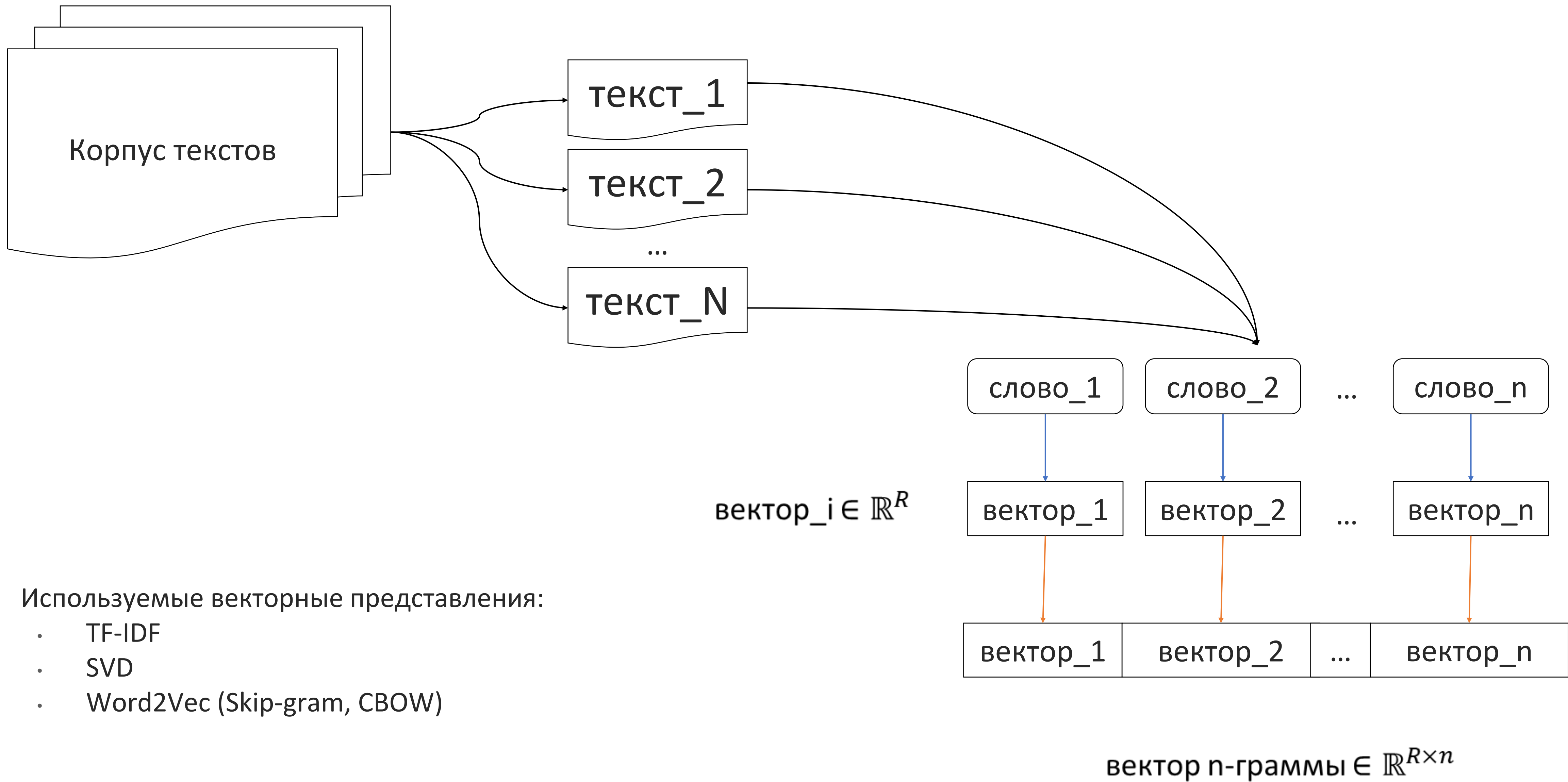
Аналитический стэк на базе реляционного OLAP создает проблему «последней мили» доставки данных до потребителей



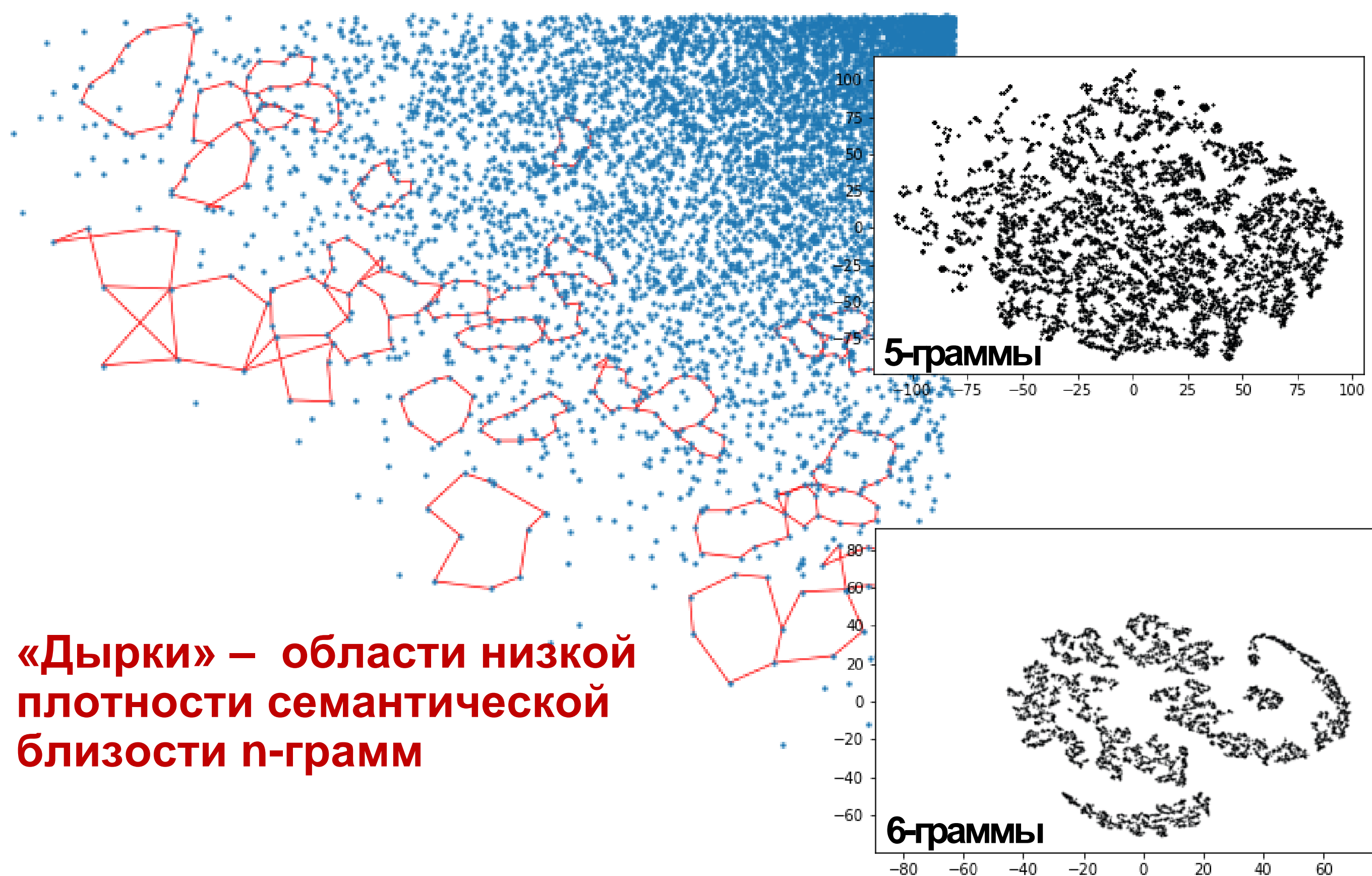
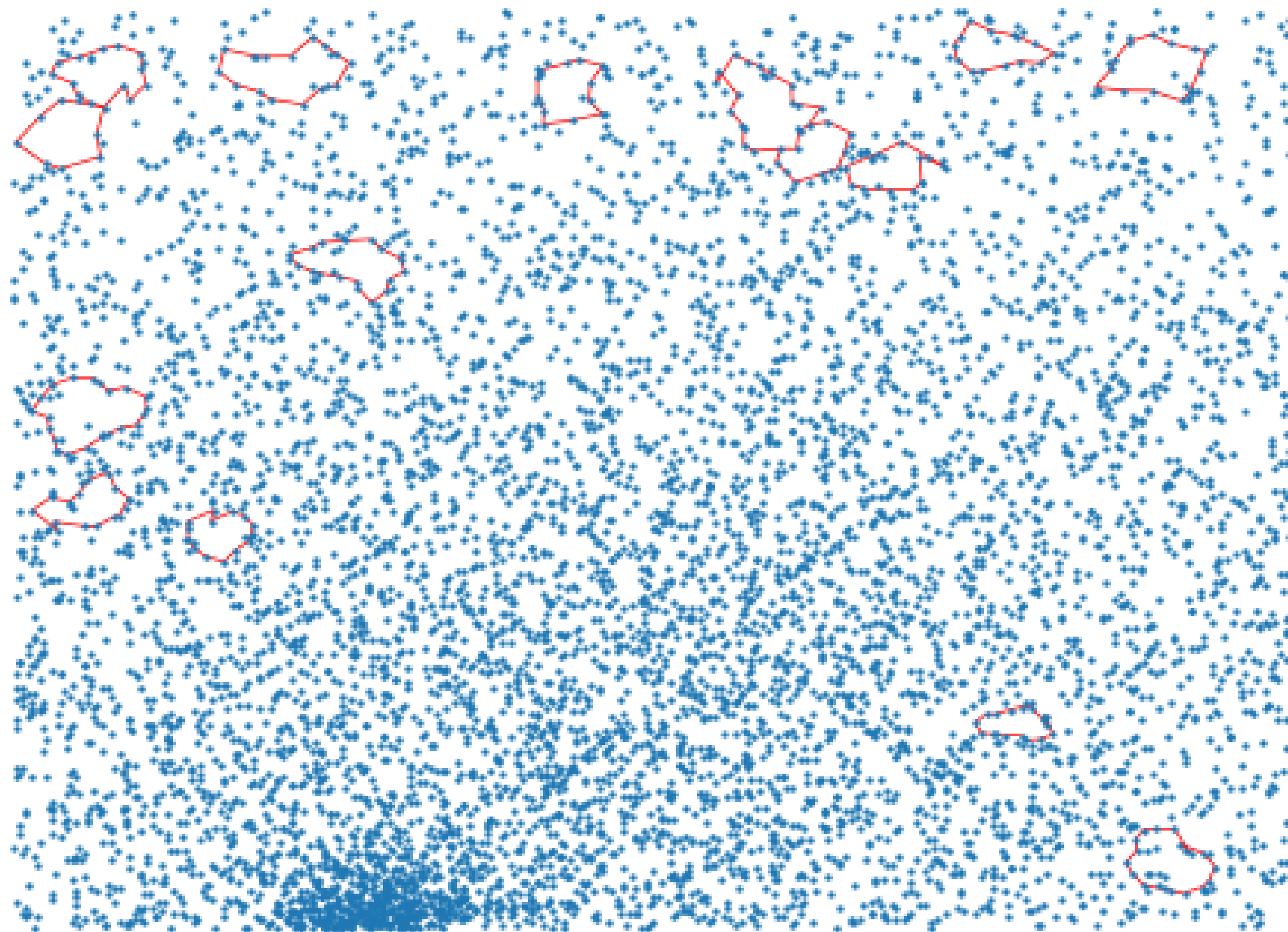
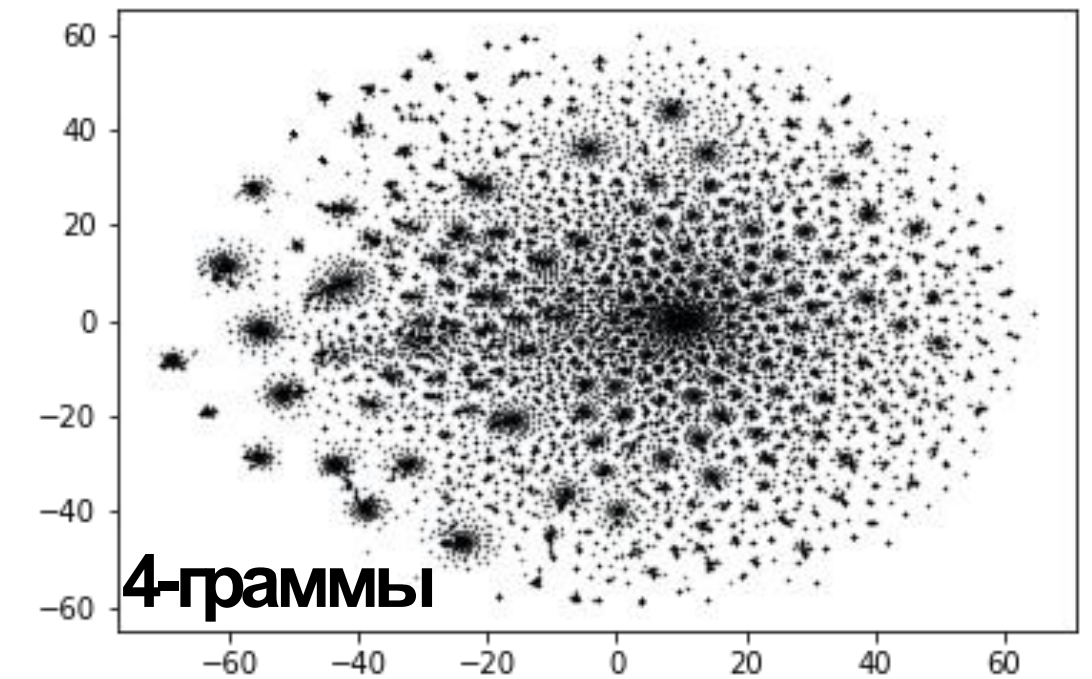
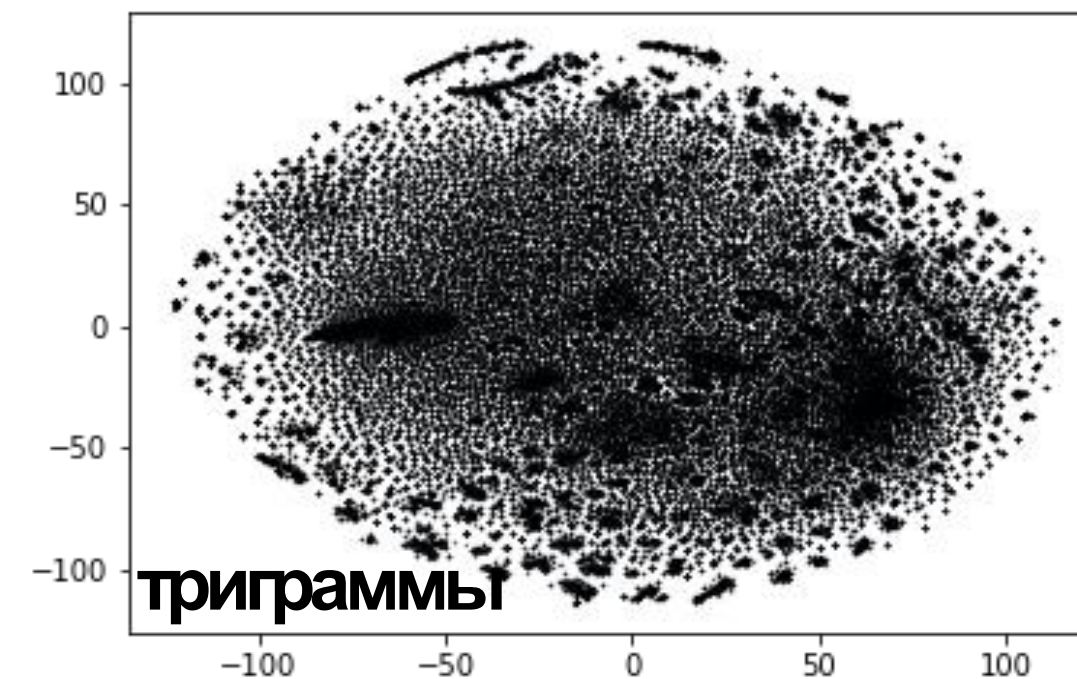
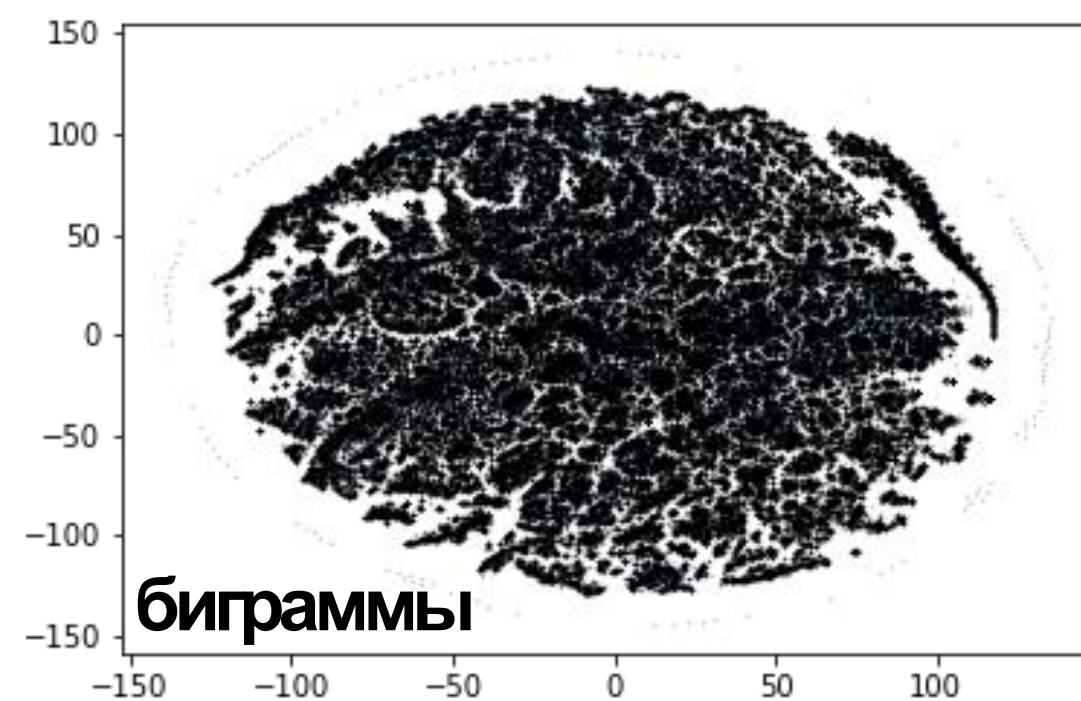
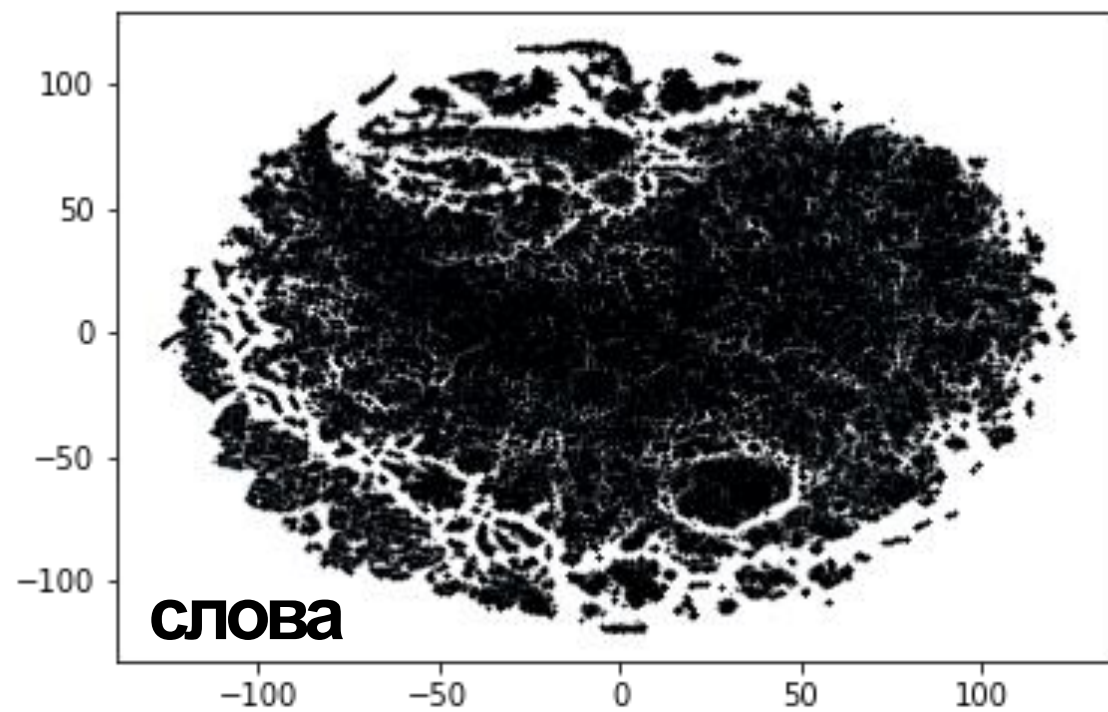
При формировании детального слоя хранилища данных предполагается его дальнейшее использование во всевозможных аналитических сценариях, а не только под конкретные бизнес-требования одного или нескольких проектов.

Но в презентационном слое хранилища витрины данных реализуются на основе текущих потребностей бизнес-пользователей - то есть на основе того, что они хотят увидеть в отчётах и дашбордах здесь и сейчас (на момент написания требований).

При этом потребности бизнес-пользователей быстро меняются, и трудозатраты на развитие и поддержку витрин кратно растут. Data-инженеры не успевают за изменяющимися потребностями потребителей данных, возникает большой бэклог по подготовке витрин. Это может привести к хаосу в работе с данными на уровне организации.



Визуальная интерпретация семантической близости n-грамм



Эмпирически обнаруженные свойства социума – концепты, в виде ограничительных свойств модели социума:

1) Социальный граф социума (социокультурное тело с общей проблемой видового выживания) представляет собой граф с количеством вершин в несколько миллиардов ($n \geq 8 \cdot 10^9$). При большой величине социальный граф связный, имеет одну гигантскую компоненту связности. Для самого большого изученного подграфа социума как соцсеть Facebook гигантская компонента составляет 99,91%;

2) Распределение вершин социального графа подчиняется степенному закону;

3) Среднее число связей между двумя случайными вершинами в социальном графе колеблется от 2 до 10, а среднее его значение от 4,4 до 6,6 что позволяет сделать предположение о маленьком диаметре социального графа: $10 \leq D \leq 10^2$;

4) Социальный граф сильно разрежен (является примером вырождения полного графа). При n -вершин, в социальном графе $k \cdot n$ -ребер, где k - константа. При этом теоретически количество ребер в полном графе равно $n \cdot (n-1)/2$;

5) Социальный граф имеет большой коэффициент кластеризации;

Теоретико-игровой подход обогатил список ограничений модели еще тремя теоретически предопределёнными, но экспериментально не обоснованными свойствами:

6) Социальный граф ориентированный;

7) Вероятность образования новой связи между двумя случайными вершинами социального графа прямо зависит от количества совпадающих клик, в которые входят эти вершины;

8) Вероятность изменения связей вершины социального графа прямо зависит от количества изменений связей в ее ближайшем социальном окружении.

Классификация подходов мат. моделирования строения социума

	Бесструктурные модели (модели на множестве N вершин)	Структурные модели	
		Не ориентированные	Ориентированные
Теоретико-вероятностный подход (связи образуются исходя из вероятностей)	Модель Эрдэша-Ранье: - маленький коэффициент кластеризации. Модель предпочтительного присоединения Барабаши-Альберт: - маленький коэффициент кластеризации. Модель Боллобаши-Риордан: - нет степенного распределения вершин; - маленький коэффициент кластеризации. Модель Бакли-Остгуса: - маленький коэффициент кластеризации. Модель Гречникова: - небольшое отклонение в степенном распределении вершин	Модель Уоттса-Строгатца (Модель с круговым графом, имитирующая линковые кольца спама) - нарушает принцип связности; - нет степенного распределения вершин.	Реляционно-иерархическая модель со строго ориентированным социальным графом Гарина - подразумевает полную перестройку графа для моделирования социальных лифтов, что противоречит наблюдениям.
Теоретико-игровой подход (связи образуются исходя из условий выгоды)	Кооперативная модель Джексона-Волински: - нет степенного распределения вершин. Модель Шелинга: - нарушает принцип связности; - нет степенного распределения вершин.	Модель Чайлз-Боргс с равновесием Нэша (переложение модели Грановеттера с одной группы на весовую совокупность приглашений нескольких групп): - не отражает феномен рекламной глухоты.	Модель Грановеттера (Модель стоячих аплодисментов) - не отражает феномен рекламной глухоты. 2.Модель социокультурного тела Тихонова (идентификационный подход) - функциональная модель не имеет математического представления
Смешанный подход	Модель копирования: - нет степенного распределения вершин, вершины имеют фиксированную валентность; - старые вершины имеют большую валентность, чем новые вершины, что противоречит наблюдениям	Эпидемиологическая модель (на базе модели Уоттса-Строгатца) - жизненный цикл вершин подчинен гипотезе о эпидемиологической природе прерывания связей	Идентификационная модель с обратной связью Гарина - не противоречит ограничениям моделирования. - рассчитана вероятность успешной передачи информации в среднестатистической коммуникации.

Тепловая карта сочетания свойств моделирования строения социума

№	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Свойство	Модель Эрдэша-Ранье	Модель предпочтительного присоединения Барабаши-Альберт	Модель Боллобаши-Риордан	Модель Бакли-Остгуса	Модель Гречникова	Кооперативная модель Джексона-Волински	Модель Шелинга	Модель копирования	Модель Уоттса-Строгатца	Модель Чайлз-Боргс с равновесием Нэша	Эпидемиологическая модель	Реляционно-иерархическая модель со строго ориентированным социальным графом Гарина	Модель стоячих аплодисментов Грановеттера	Модель социального тела проблемы Тихонова	Реляционно-иерархическая модель с обратной связью Гарина
1	+	+	+	+	+	+	-	+	-	+	+	+	+	+	+
2	+	+	-	+	-	-	-	-	-	-	+	+	-	-	+
3	-	-	+	-	+	+	+	+	+	+	-	-	+	-	-
4	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+
5	-	-	-	-	+	+	+	+	+	+	+	+	+	+	+
6	-	-	+	+	-	-	+	-	-	-	+	+	+	+	+
7	-	+	-	-	-	+	+	+	+	+	+	-	+	+	+
8	-	+	-	-	-	+	+	-	-	+	+	-	+	+	+

Процесс поиска дырок в концептах свойств моделирования социума

1									2									3									4									5													
	1	2	3	4	5	6	7	8																																									
1	0	1	0	1	0	0	0	0		1	0	1	0	1	0	0	1	1		1	0	0	1	1	0	1	0	0		1	0	0	1	1	1	0	0	0		1	0	0	1	1	0	0	0		
2	1	0	0	1	0	0	0	0		2	1	0	0	1	0	0	1	1		2	0	0	0	0	0	0	0	0		2	1	0	0	1	0	1	0	0		2	0	0	0	0	0	0	0	0	
3	0	0	0	0	0	0	0	0		3	0	0	0	0	0	0	0	0		3	1	0	0	1	0	1	0	0		3	0	0	0	0	0	0	0	0		3	1	0	0	1	1	0	0	0	
4	1	1	0	0	0	0	0	0		4	1	1	0	0	0	0	1	1		4	1	0	1	0	0	1	0	0		4	1	1	0	0	0	1	0	0		4	1	0	1	0	1	0	0	0	
5	0	0	0	0	0	0	0	0		5	0	0	0	0	0	0	0	0		5	0	0	0	0	0	0	0	0		5	0	0	0	0	0	0	0	0		5	1	0	1	1	0	0	0	0	
6	0	0	0	0	0	0	0	0		6	0	0	0	0	0	0	0	0		6	1	0	1	1	0	0	0	0		6	1	1	0	1	0	0	0	0		6	0	0	0	0	0	0	0	0	
7	0	0	0	0	0	0	0	0		7	1	0	0	1	0	0	0	0	1		7	0	0	0	0	0	0	0	0		7	0	0	0	0	0	0	0	0		7	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0		8	1	1	0	1	0	0	1	0		8	0	0	0	0	0	0	0	0		8	0	0	0	0	0	0	0	0		8	0	0	0	0	0	0	0	0	
6									7									8									9									10													
	1	2	3	4	5	6	7	8																																									
1	0	0	1	1	1	0	1	1		1	0	0	0	0	0	0	0	0		1	0	0	1	1	1	0	1	0		1	0	0	0	0	0	0	0	0		1	0	0	1	1	1	0	1	1	
2	0	0	0	0	0	0	0	0		2	0	0	0	0	0	0	0	0		2	0	0	0	0	0	0	0	0		2	0	0	0	0	0	0	0	0		2	0	0	0	0	0	0	0	0	
3	1	0	0	1	1	0	1	1		3	0	0	0	1	1	1	1	1		3	1	0	0	1	1	0	1	0		3	0	0	0	1	1	1	0	0		3	1	0	0	1	1	0	1	1	
4	1	0	1	0	1	0	1	1		4	0	0	1	0	1	1	1	1		4	1	0	1	0	1	0	1	0		4	0	0	1	0	1	1	0	0		4	1	0	1	0	1	0	1	1	
5	1	0	1	1	0	0	1	1		5	0	0	1	1	0	1	1	1		5	1	0	1	1	0	0	1	0		5	0	0	1	1	0	1	0	0		5	1	0	1	1	0	0	1	1	
6	0	0	0	0	0	0	0	0		6	0	0	1	1	1	0	1	1		6	0	0	0	0	0	0	0	0		6	0	0	1	1	1	0	0	0		6	0	0	0	0	0	0	0	0	
7	1	0	1	1	1	0	0	1		7	0	0	1	1	1	1	0	1		7	1	0	1	1	1	0	0	0		7	0	0	0	0	0	0	0	0		7	1	0	1	1	1	0	0	1	
8	1	0	1	1	1	0	1	0		8	0	0	1	1	1	1	1	0		8	0	0	0	0	0	0	0	0		8	0	0	0	0	0	0	0	0		8	1	0	1	1	1	0	1	0	
11									12									13									14									1 - 14													
	1	2	3	4	5	6	7	8																																									
1	0	1	0	1	1	1	1	1		1	0	1	0	1	1	1	0	0		1	0	0	1	1	1	1	1		1	0	0	0	0	1	1	1	1		1	0	5	6	11	8	6	7	6		
2	1	0	0	1	1	1	0	1		2	1	0	0	1	1	1	0	0		2	0	0	0	0	0	0	0	0		2	0	0	0	0	0	0	0	0		2	5	0	0	5	2	3	1	2	
3	0	0	0	0	0	0	0	0		3	0	0	0	0	0	0	0	0		3	1	0	0	1	1	1	1		3	0	0	0	0	0	0	0	0		3	6	0	0	8	7	4	5	4		
4	1	1	0	0	1	1	1	1		4	1	1	0	0	1	1	0	0		4	1	0	1	0	1	1	1		4	0	0	0	0	0	0	0	0		4	11	5	8	0	9	7	7	6		
5	1	1	0	1	0	1	1	1		5	1	1	0	1	0	1	0	0		5	1	0	1	1	0	1	1		5	1	0	0	0	0	1	1	1		5	8	2	7	9	0	6	7	6		
6	1	1	0	1	1	0	1	1		6	1	1	0	1	1	0	0	0		6	1	0	1	1	1	0	1		6	1	0	0	0	1	0	1	1		6	6	3	4	7	6	0	4	4		
7	1	1	0	1	1	1	0	1		7	0	0	0	0	0	0	0	0		7	1	0	1	1	1	1	0	1		7	1	0	0	0	1	1	0	1		7	7	1	5	7	7	4	0	7	
8	1	1	0	1	1	1	1	0		8	0	0	0	0	0	0	0	0		8	1	0	1	1	1	1	1	0		8	1	0	0	0	1	1	1	0		8	6	2	4	6	6	4	7	0	

Экспериментально подтвержденные свойства в ходе диссертационного исследования:

1) Социальный граф социокультурного тела является связным. Показатель связности в социокультурном теле эксперимента – 74,77%.

2) Распределение степеней вершин социального граф социокультурного тела подчиняется степенному закону.

3) Среднее число связей между двумя вершинами в успешных цепочках передачи данных в экспериментах по распространению информации в социокультурном теле колеблется от 2 до 12. Этого факта не достаточно что бы сделать окончательный выводы о средней величине связей между случайными вершинами в социальном графе и о величине диаметра социального графа, из-за большого количества прерываний цепочек передачи данных – до 65% прерванных цепочек передачи данных на каждом этапе передачи данных.

4) Социальный граф социокультурного тела сильно разрежен (является примером вырождения полного графа). При n -вершин, в социальном графе $k \cdot n$ -ребер, где k - константа. Теоретически количество ребер в полном графе равно $\approx n^2/2$. В эксперименте в Facebook данная константа вычислена близкой к $k \approx 2$.

5) Социальный граф социокультурного тела имеет большой коэффициент кластеризации. Значение коэффициента кластеризации для социального графа социокультурного тела экспериментально определено равным 0,0162.

б) Социальный граф социокультурного тела строго ориентированный. Это свойство в рамках данного диссертационного исследования доказано в результате детального анализа результатов эксперимента Уоттса.

7) Вероятность образования новой связи между двумя случайными вершинами социального графа прямо зависит от количества совпадающих клик, в которые входят эти вершины. Косвенно на это указывают найденная в диссертационном исследовании корреляция между показателем коэффициента кластеризации и степенью вершины социального графа, входящего в кластер.

8) Вероятность изменения связей вершины социального графа прямо зависит от количества изменений связей в ее ближайшем социальном окружении. Косвенно это свойство доказано в рамках диссертационного исследования через введение нового понятия порога перколяции.

9) Социокультурное тело только частично пропускает через себя проходящую информацию. Это свойство к частичному пропуску - «просачиванию» информации получило наименование перколяции.

Перколяция измеряется через показатель «коэффициента перколяции» (K_{perc}) – отношение количества исходящих каналов связи к входящим каналам связи. Коэффициент перколяции постепенно снижается на каждом этапе передачи информации по линейному закону, но в очень маленьком диапазоне. Усредненное значение коэффициента перколяции экспериментально определено по результатам эксперимента Уотса: $K_{perc\sum} = 0,3519 \pm 0,03$.

10) При многократном прохождении через социокультурное тело однородной информации социокультурное тело теряет способность к распространению информации (Эффект рекламной глухоты). Предельное количество повторов информации, при котором социокультурное тело теряет способность к распространению информации, получило обозначение «предела перколяции» – $\max \text{Limitperc}$. Для каждого слоя индивидуума в отдельности в социокультурном теле предел перколяции может быть рассчитан индивидуально. При этом экспериментально обнаружено, что эффективность распространения информации в социокультурном теле в зависимости от количества повторов является степенной зависимостью, и рассчитывается по формуле: $y \approx 0,02 \cdot N \cdot x^{-1,2}$, где y – количество людей, воспринявших информацию (рекламу) и передавшую ее другим людям, N – размер аудитории; x – количество повторов. При количестве повторов $x \approx 30$ ($\max \text{Limitperc} \approx 30$) в зависимости от величины N эффективность распространения информации в социокультурном теле стремиться к 0, т.е. при достижении предела перколяции социокультурное тело теряет способность к распространению информации. Данная модель расчета эффективности распространения информации в социокультурном теле является объяснительной моделью эмпирически обнаруженного «эффекта рекламной глухоты».

11) Из-за высокой информационной загруженности социокультурное тело не прозрачно для любой информации. Что бы преодолеть сопротивление к распространению информации часто требуется повторение однородной информации. Такое многократное повторение информации дает возможность преодолеть сопротивление социокультурного тела. Свойство социокультурного тела приобретать способность к распространению информации при многократном прохождении однородной информации получило обозначение «порога перколяции» – $\min \text{Limitperc}$. Порог перколяции ($\min \text{Limitperc}$) – это минимальное значение исходящих каналов информации (либо частоты повтора информации), передача по которым однотипной информации приводит к возникновению у социокультурного тела способности к распространению информации ($\text{Kperc} > 0$). Для каждого отдельного индивидуума в социокультурном теле значение порога перколяции является индивидуальной характеристикой. Экспериментально обнаружена линейная зависимость между количеством связей (степенью вершин) и порогом перколяции: $y = 0,9533x + 0,8807$ (2.4.1), где y – среднее количество связей (степеней вершины), x – значение порога перколяции ($\min \text{Limitperc}$). Среднее значение порога перколяции обнаруженное экспериментально $\min \text{Limitperc} \sum \approx 3$ (При этом для 42% индивидуумов в социокультурном теле эксперимента имеют порог перколяции $\min \text{Limitperc} \sum = 1$)