

Проверка возможности автоматического сбора релевантных экспериментальных данных

AI4Science центр, Сбер

Идея



Получение представления о научной области



Получение представления о научной области

Как обстоят дела с исследованиями в этой экспериментальной области?



Получение представления о научной области

Как обстоят дела с исследованиями в этой экспериментальной области?



Вот недавний детальный метаобзор области от уважаемых авторов в солидном рецензируемом журнале.

В нём авторы

- * исследовали сотни работ по теме,
- * представили ссылки на оригинальные исследования
- * качественно обработали результаты



Получение представления о научной области

Как обстоят дела с исследованиями в этой экспериментальной области?



Вот недавний детальный метаобзор области от уважаемых авторов в солидном рецензируемом журнале.

В нём авторы

- * исследовали сотни работ по теме,
- * представили ссылки на оригинальные исследования
- * качественно обработали результаты



Такие метаобзоры есть не всегда

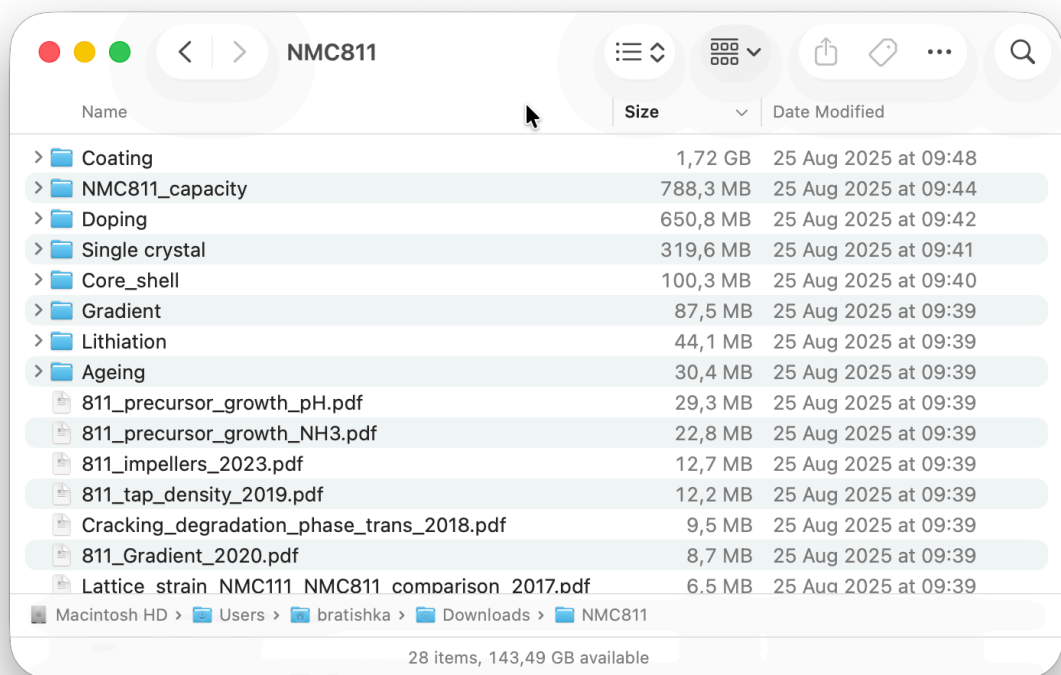
Составление метаобзора области на примере материала NMC811 для катодов в Li-Ion аккумуляторах

Какие практические свойства будет иметь аккумулятор на основе материала $\text{LiNi}_{0.8}\text{Mn}_{0.1}\text{Co}_{0.1}\text{O}_2$ для катода?



Составление метаобзора области на примере материала NMC811 для катодов в Li-Ion аккумуляторах

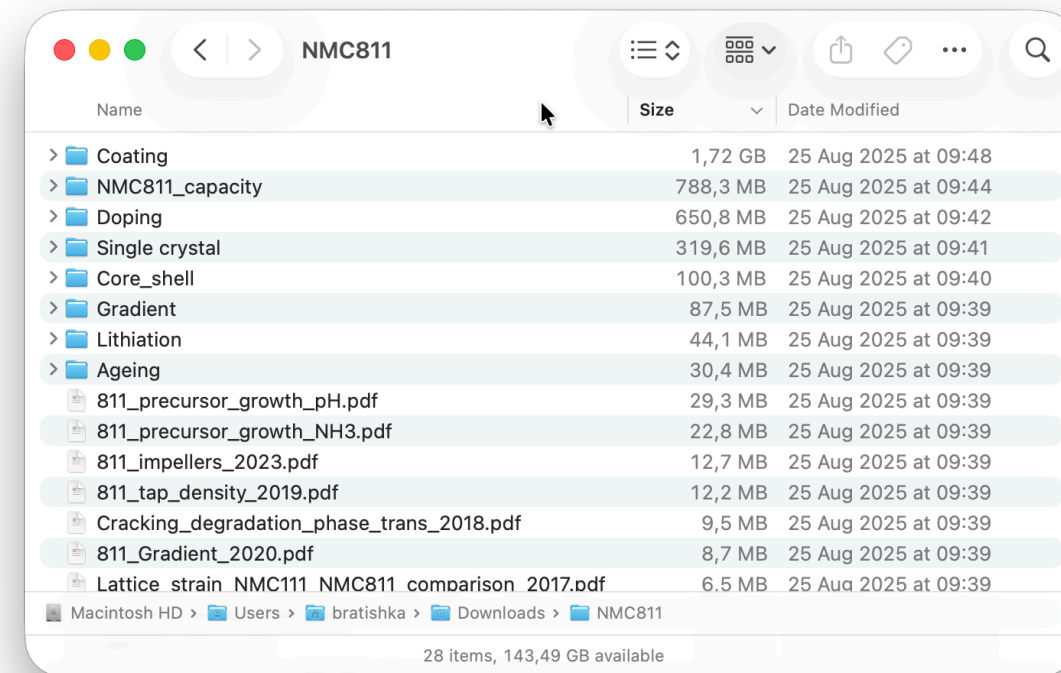
Поиск и скачивание
700+ научных статей



Составление метаобзора области на примере материала NMC811 для катодов в Li-Ion аккумуляторах

Поиск и скачивание 700+ научных статей

Поиск значений релевантных характеристик в каждой статье

[illegible]

Составление метаобзора области на примере материала NMC811 для катодов в Li-Ion аккумуляторах

Поиск и скачивание 700+ научных статей

Поиск значений релевантных характеристик в каждой статье

Обработка результатов и написание текста

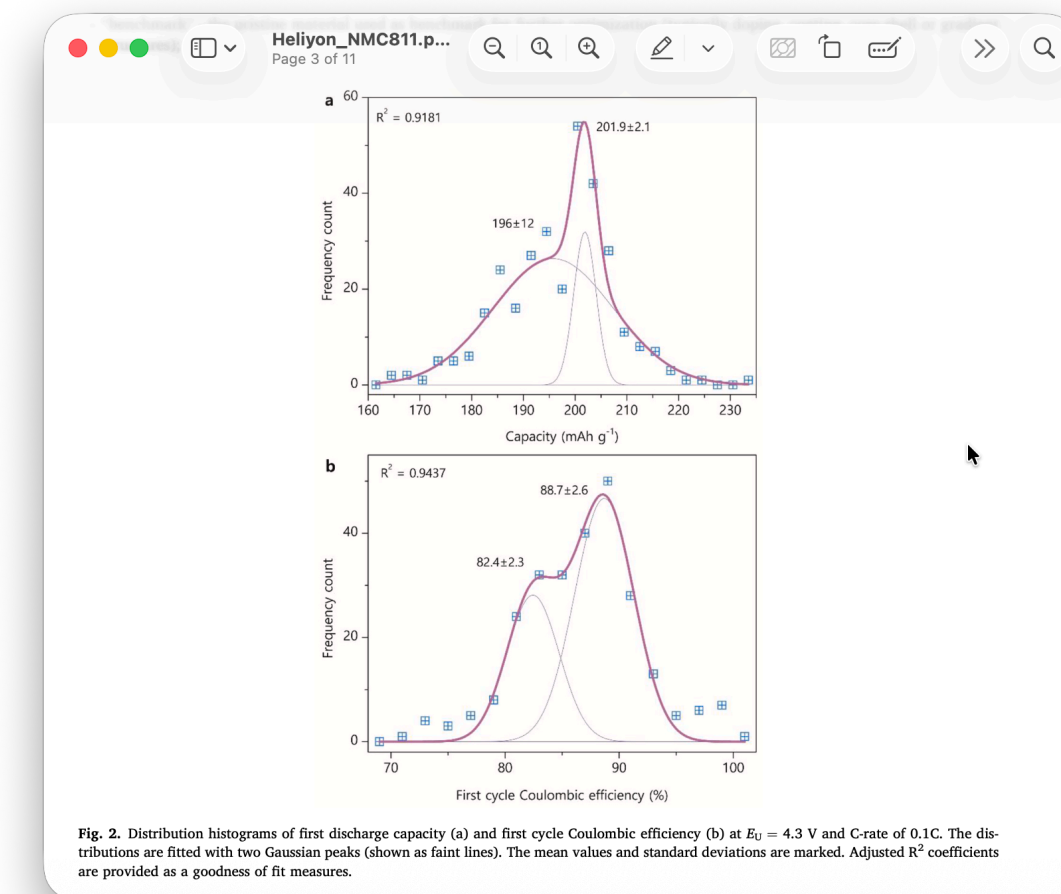


+

Sheet1

Table for publication 2

The raw numerical data extracted from literature and Digital Object Identifiers (DOI) of the source publications.	Composition	First charge/discharge cycle at room temperature	Voltage comp. V	Initial coulombic efficiency %	Capacity 1 st cycle mAh/g	Rate	Upper voltage, V	# of cycles	Capacity retention %	Rate capacity at 4.3V upper voltage, room temperature	No content in this work	Cathode composition, weight %	Binder	DOI or web reference	Tags			
	Composition	First discharge cycle capacity, mAh/g	Voltage comp. V	Initial coulombic efficiency %	Capacity 1 st cycle mAh/g	Rate	Upper voltage, V	# of cycles	Capacity retention %	Capacity fade 1 st cycle	No content in this work	Active material	Conductive additive	Binder	DOI or web reference	Tags		
	UNIM-BMNO-100-1002	200	2.7-4.3	85	151	0.5	4.3	100	94	0.060				90	5.5	4.5	10.3393/arc.2019.043890	polycrystalline, synthesis, lab-made
	UNIM-BMNO-100-1002	207	2.7-4.4		150	0.5	4.4	100	91	0.059				90	5.5	4.5	10.3393/arc.2019.043890	polycrystalline, synthesis, lab-made
	UNIM-BMNO-100-1002	171	2.7-4.3	90	151	0.5	4.3	100	91	0.059				90	5.5	4.5	10.3393/arc.2019.043890	polycrystalline, synthesis, lab-made
	Li O3BMNO-100-1002 (LiO3BMNO)	164	2.7-4.3	93	143	0.5					7.9						10.1002/anie.202001151	polycrystalline, commercial
	Li O3BMNO-100-1002 (LiO3BMNO)	168	2.7-4.3	93	143	0.5					7.9						10.1002/anie.202001151	polycrystalline, commercial
	UNIM-BMNO-100-1002	175	2.7-4.3		151	0.5								80	10	10	10.1039/c3cc30179a	polycrystalline, commercial
	UNIM-BMNO-100-1002	181	2.7-4.3	92	151	0.5								80	10	10	10.1039/c3cc30179a	polycrystalline, commercial
	UNIM-BMNO-100-1002	200	2.7-4.4	91	151	0.5								80	10	10	10.1039/c3cc30179a	polycrystalline, commercial
	UNIM-BMNO-100-1002	200	2.7-4.3	87	151	0.5	4.3	100	93	0.750				80	10	10	10.1039/c5cc00662	polycrystalline, commercial
	UNIM-BMNO-100-1002	198	2.7-4.4	91	151	0.5	4.3	100	91	0.750		3.88		80	10	10	10.1039/c5cc00662	polycrystalline, benchmark, lab-made
	UNIM-BMNO-100-1002	174	2.7-4.3	87	151	0.5	4.3	200	77	0.160	155	138	135	80	10	10	10.1039/c7cc31980	polycrystalline, commercial
	UNIM-BMNO-100-1002	188	2.8-4.3	90	151	0.5	4.3	200	80	0.300	100	110	112	80	10	10	10.1039/c7cc31980	polycrystalline, benchmark, commercial
	UNIM-BMNO-100-1002	200	2.8-4.3	90	151	0.5	4.3	100	94	0.400				80	10	10	10.1039/c7cc31980	polycrystalline, benchmark, commercial
	UNIM-BMNO-100-1002	188	2.8-4.3	90	150	0.5	4.3	200	80	0.300				80	10	10	10.1039/c7cc31980	polycrystalline, benchmark, commercial
	UNIM-BMNO-100-1002	179	2.8-4.2		150	0.33	4.2	375	80	0.051				95	2	2	10.1039/c7cc31980	polycrystalline, benchmark, commercial
	UNIM-BMNO-100-1002	188	2.8-4.3		150	0.5								95	2	2	10.1039/c7cc31980	polycrystalline, benchmark, commercial
	UNIM-BMNO-100-1002	200	2.7-4.3	90	151	0.5	4.3	200	90	0.400				95	2	2	10.1039/c7cc31980	polycrystalline, benchmark, commercial
	UNIM-BMNO-100-1002	200	2.7-4.3	90	151	0.5	4.3	200	88	0.360	195	187	174	118	4.4	95	10.1001/jacs.20190487	single crystal, synthesis, lab-made
	UNIM-BMNO-100-1002	185	2.7-4.2	91	151	0.33	4.2	200	87	0.060				95	3	3	10.1001/jacs.20190487	polycrystalline, synthesis, lab-made

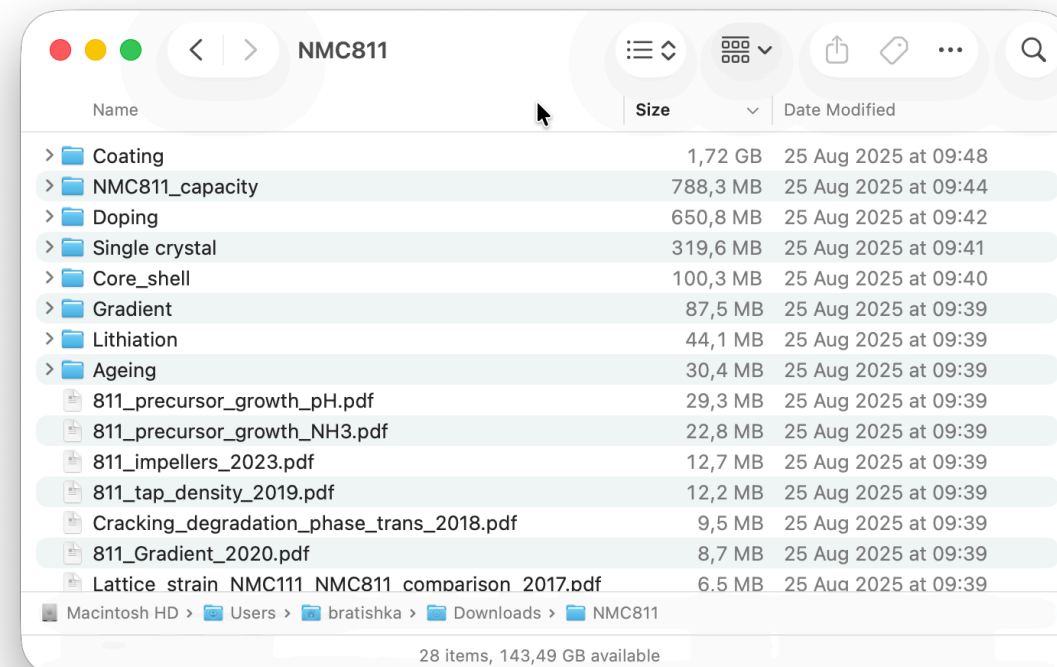


Составление метаобзора области на примере материала NMC811 для катодов в Li-Ion аккумуляторах

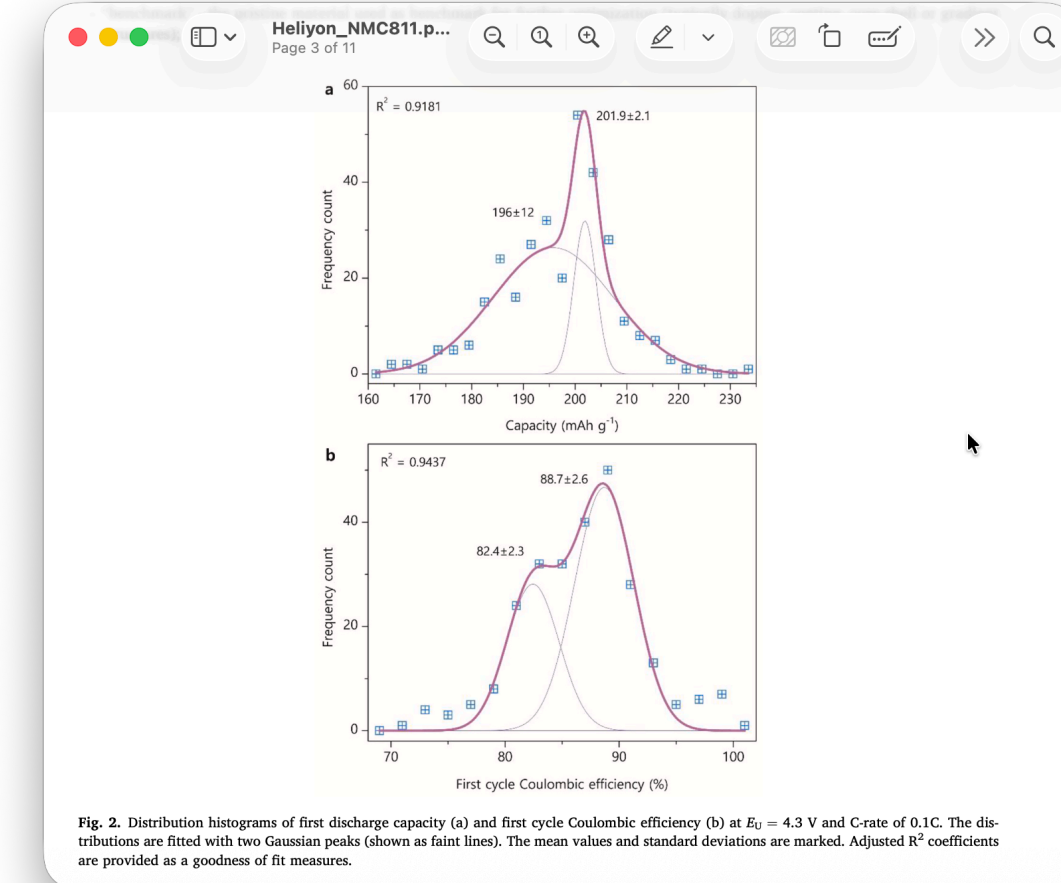
Поиск и скачивание 700+ научных статей

Поиск значений релевантных характеристик в каждой статье

Обработка результатов и написание текста



</

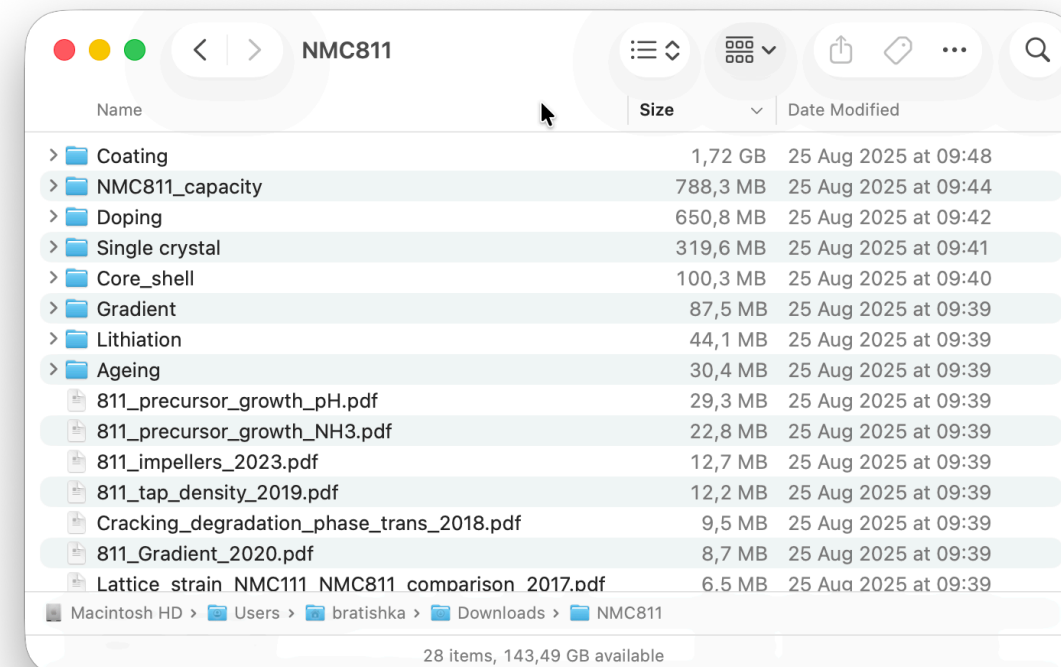
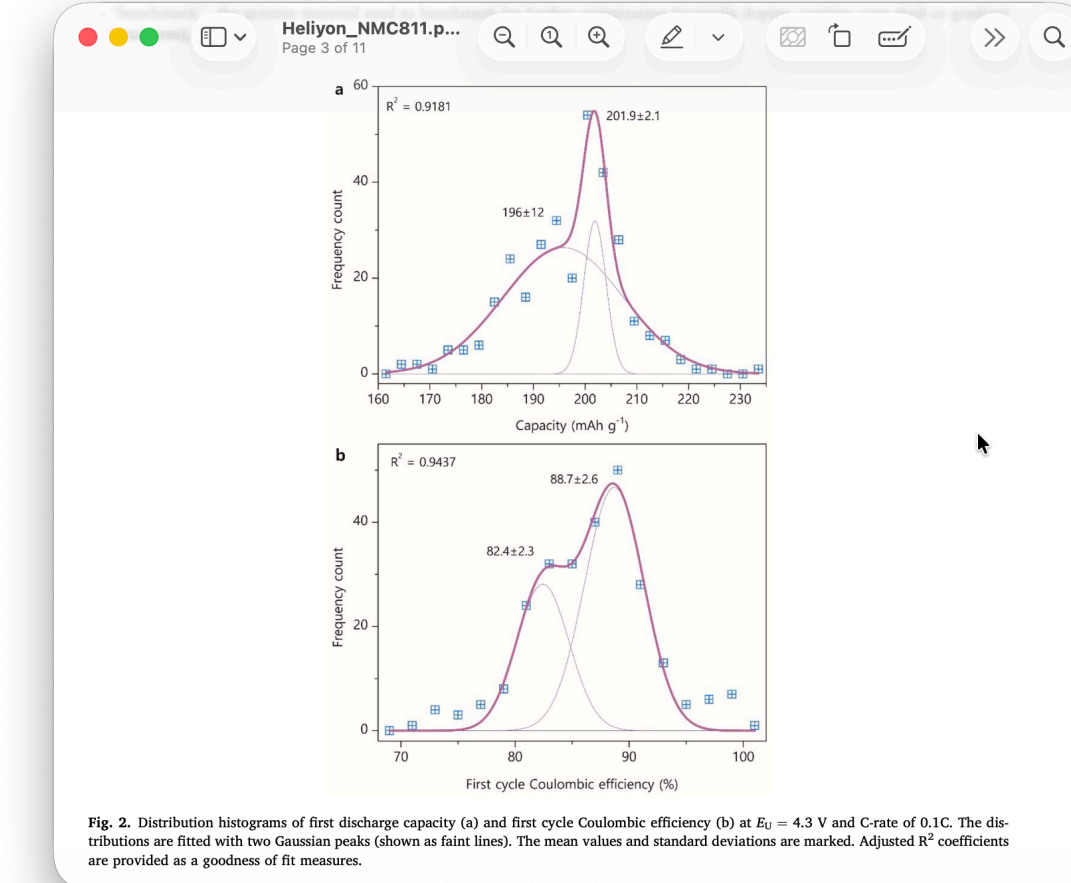


Составление метаобзора области на примере материала NMC811 для катодов в Li-Ion аккумуляторах

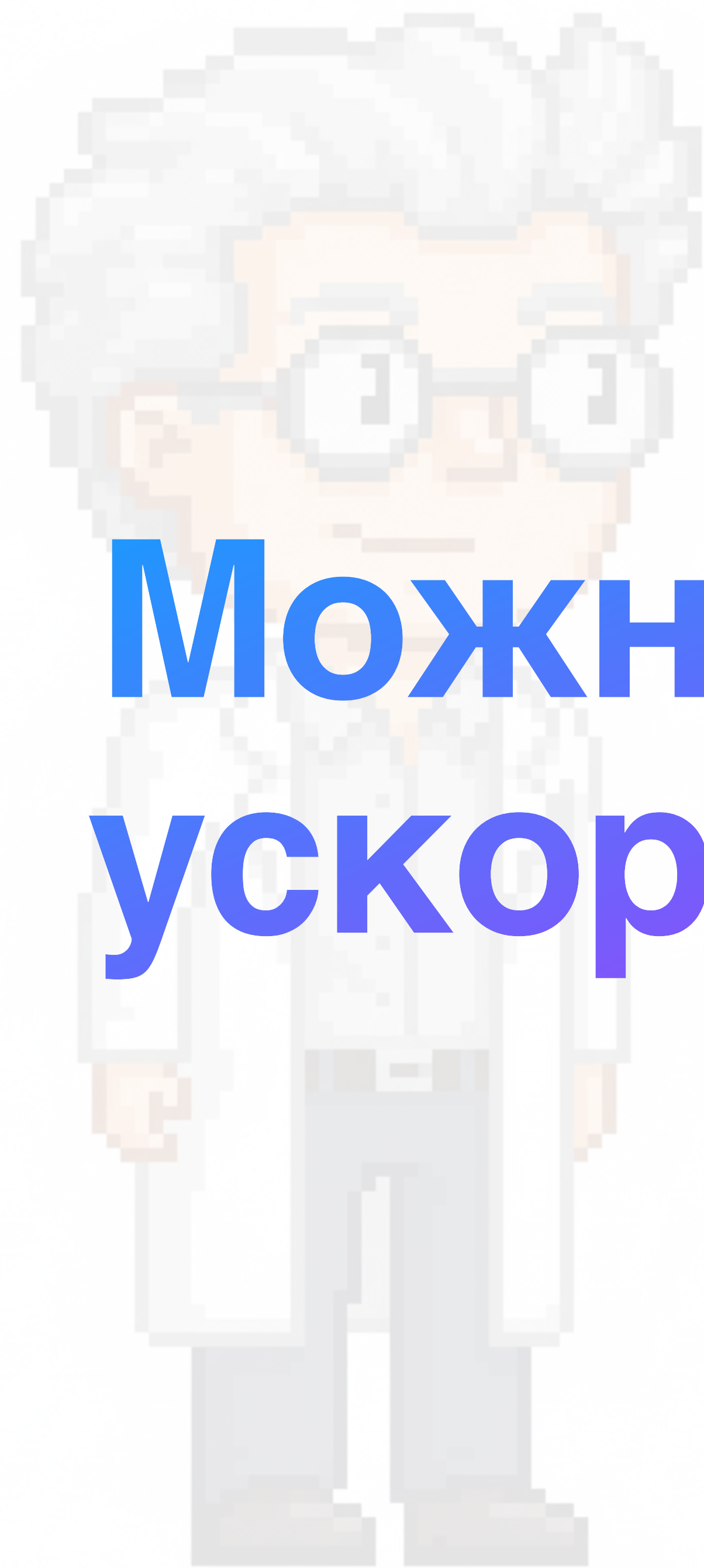
Поиск и скачивание 700+ научных статей

Поиск значений релевантных характеристик в каждой статье

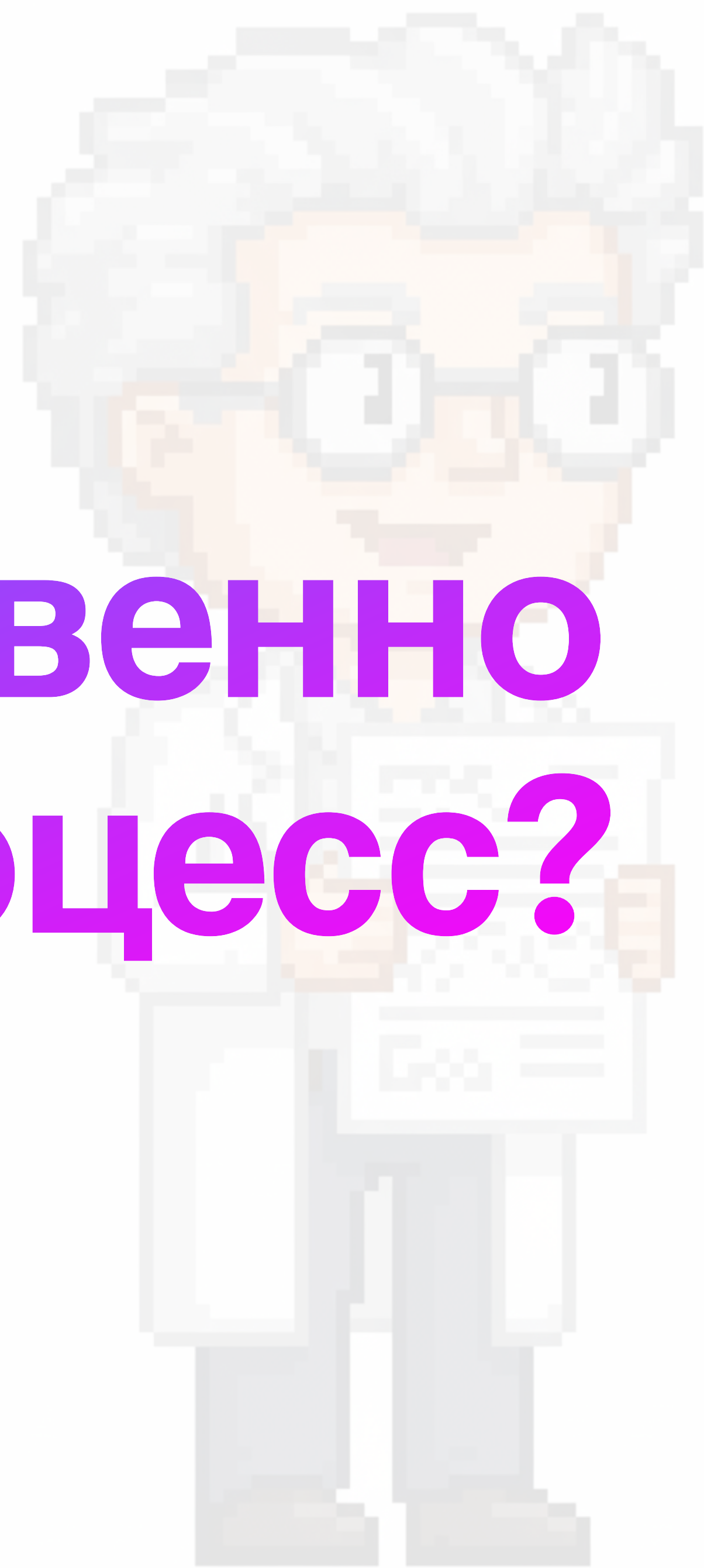
Обработка результатов и написание текста

[illegible]

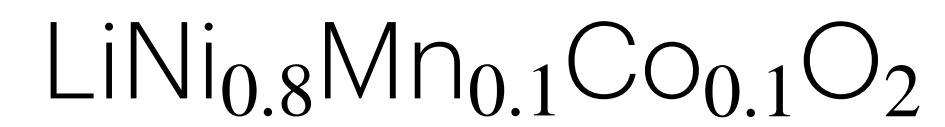
3 месяца работы!



**Можно ли существенно
ускорить этот процесс?**



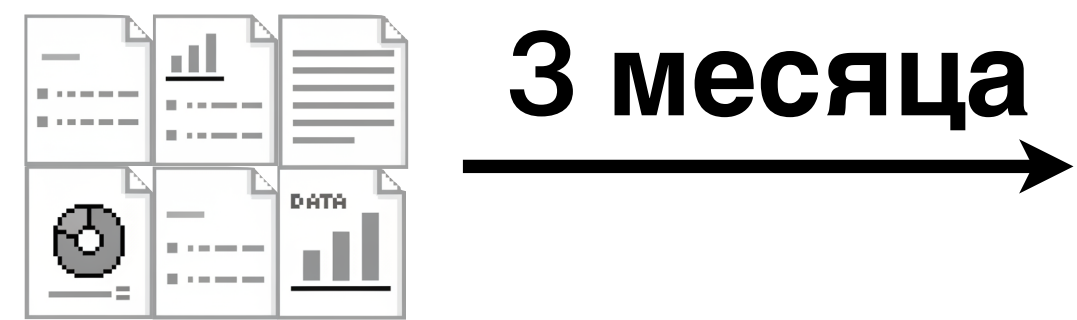
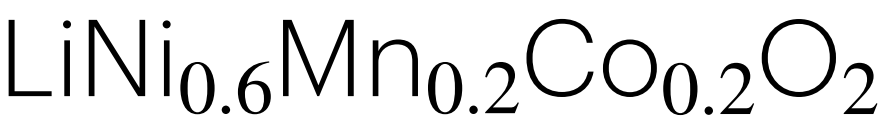
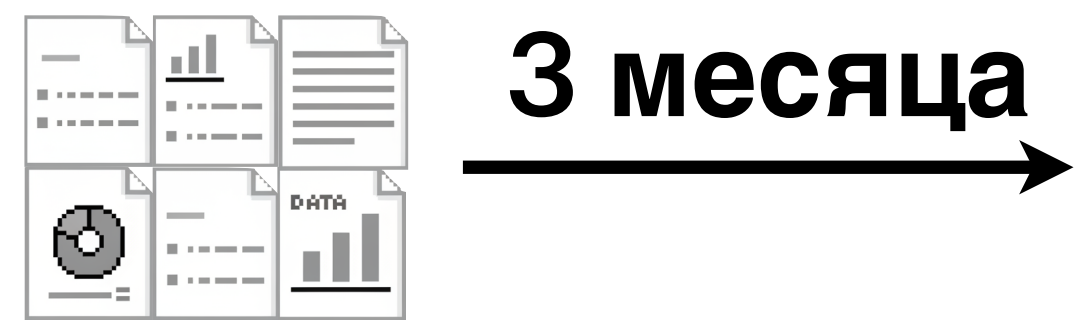
Агентное составление метаобзоров в электрохимии



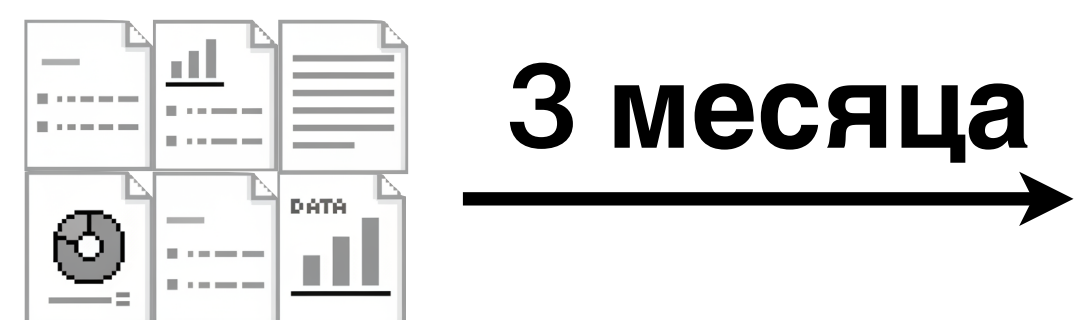
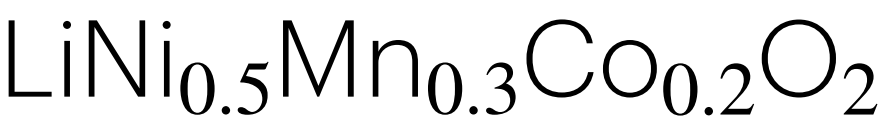
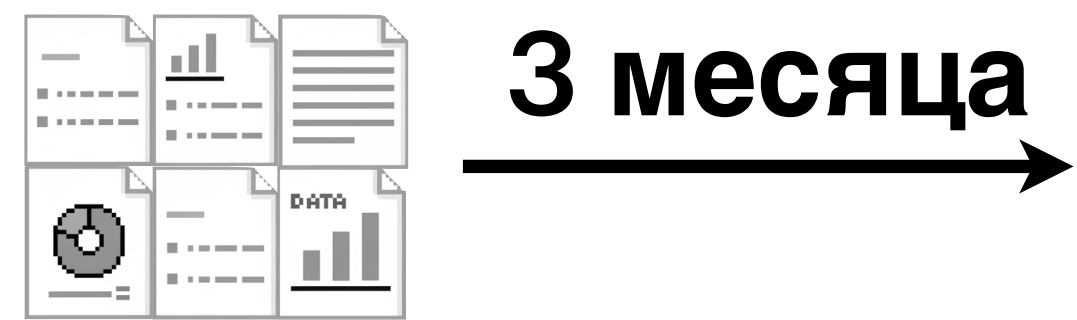
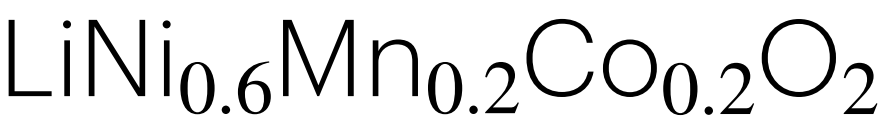
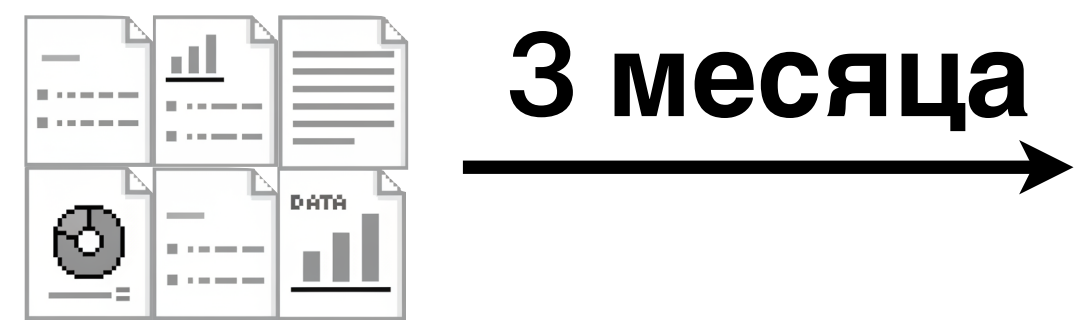
3 месяца



Агентное составление метаобзоров в электрохимии



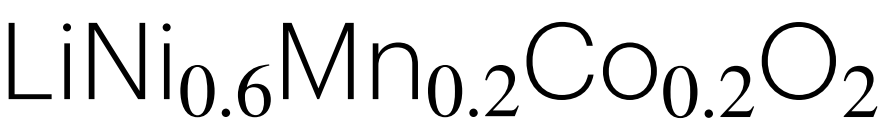
Агентное составление метаобзоров в электрохимии



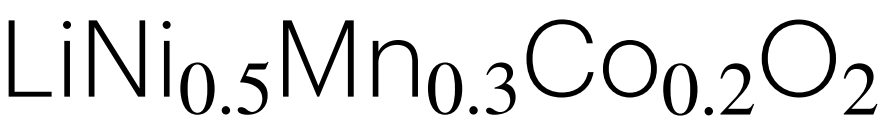
Агентное составление метаобзоров в электрохимии



3 месяца
→

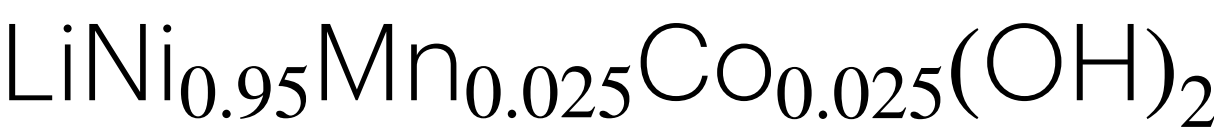


3 месяца
→



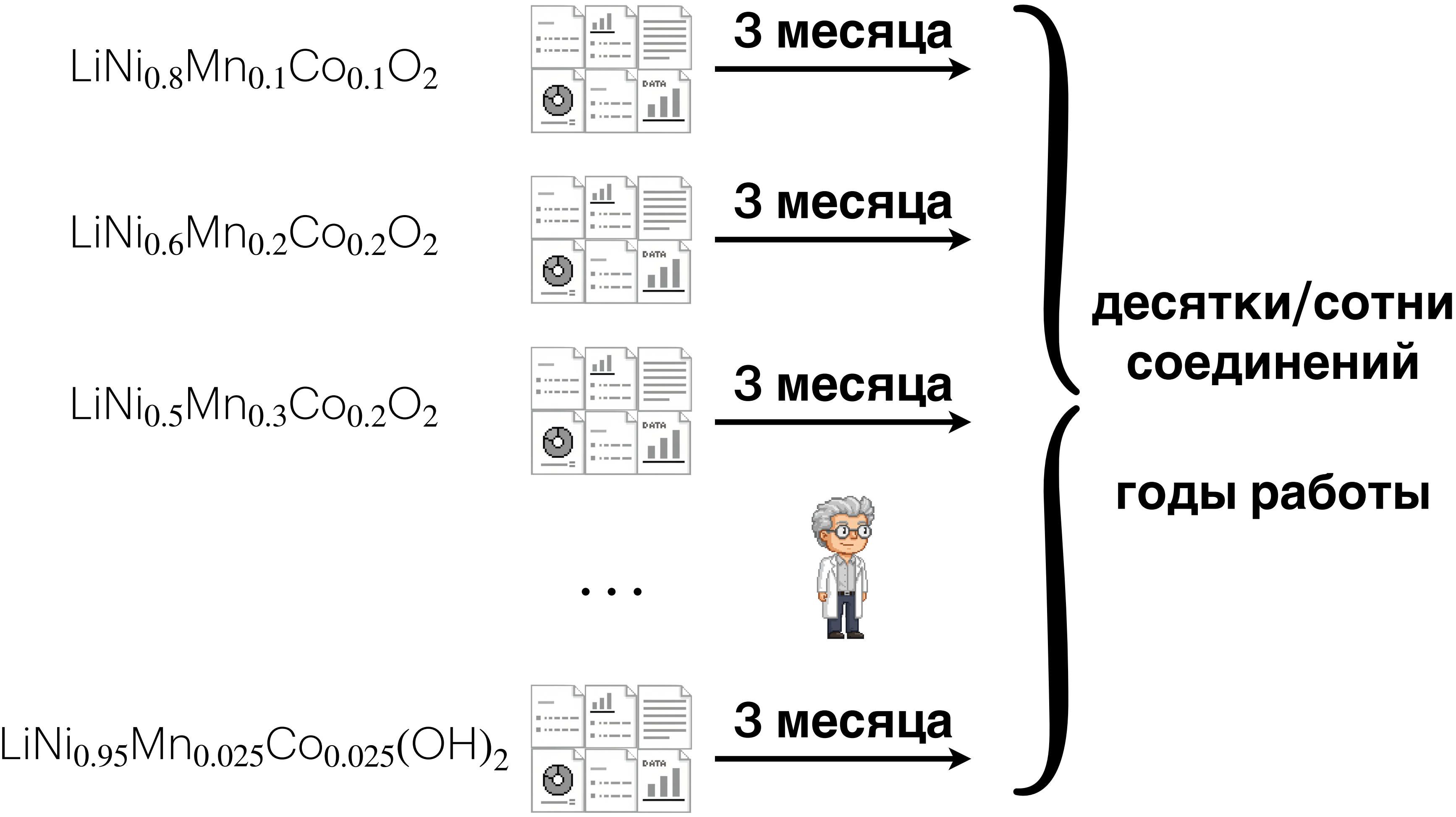
3 месяца
→

...

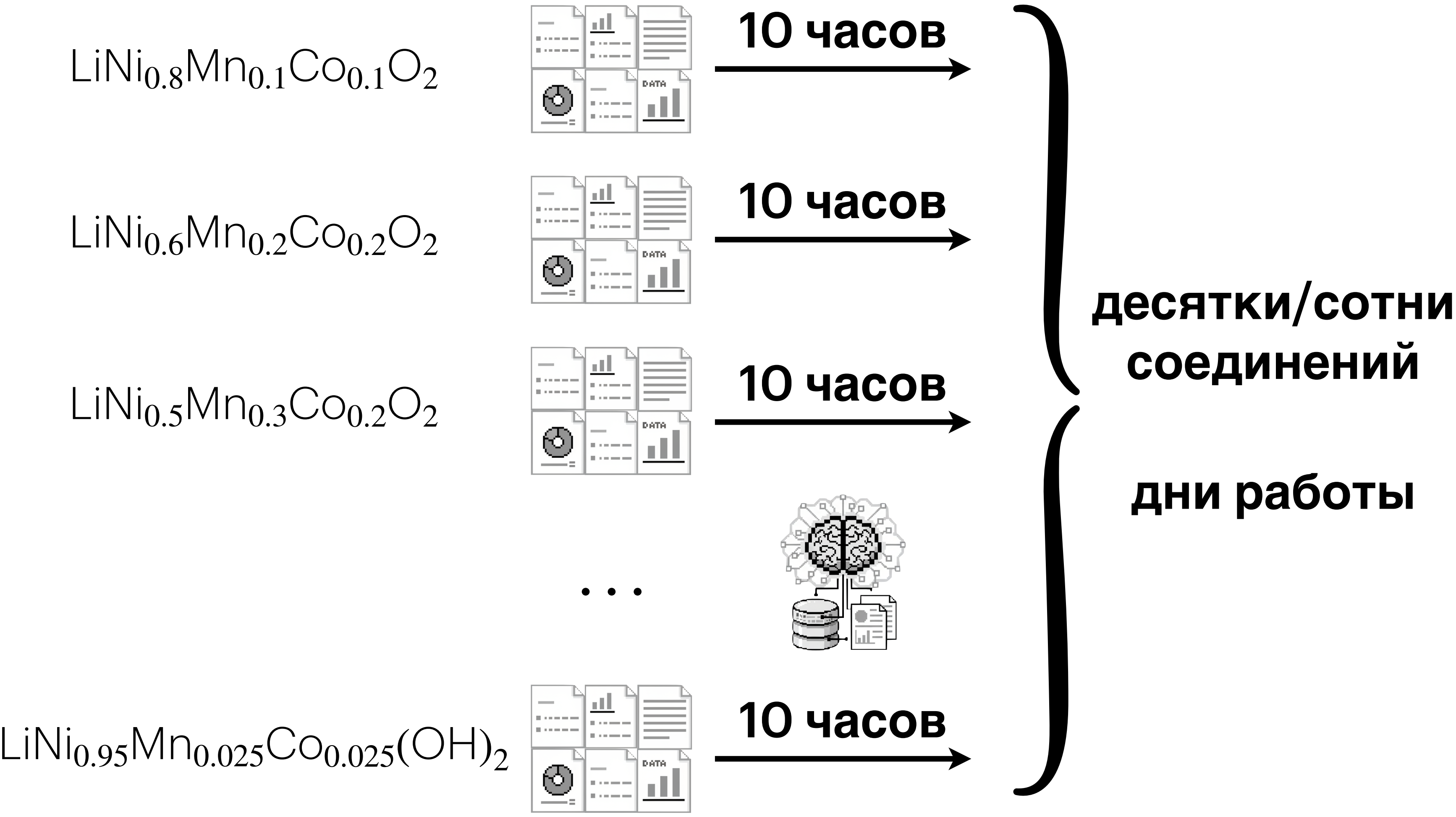


3 месяца
→

Агентное составление метаобзоров в электрохимии



Агентное составление метаобзоров в электрохимии



Агентное составление метаобзоров в электрохимии



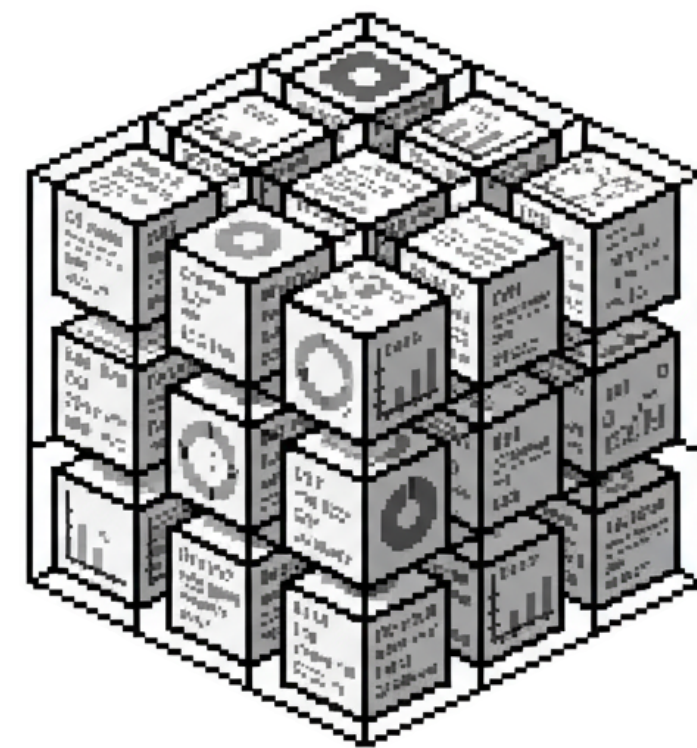
Наличие такой базы экспериментальных данных может быть полезно для дизайна новых экспериментов

Мне нужен материал для катода аккумулятора с содержанием Никеля не более 70 процентов и удельной ёмкостью не менее 198 мАч/г



Наличие такой базы экспериментальных данных может быть полезно для дизайна новых экспериментов

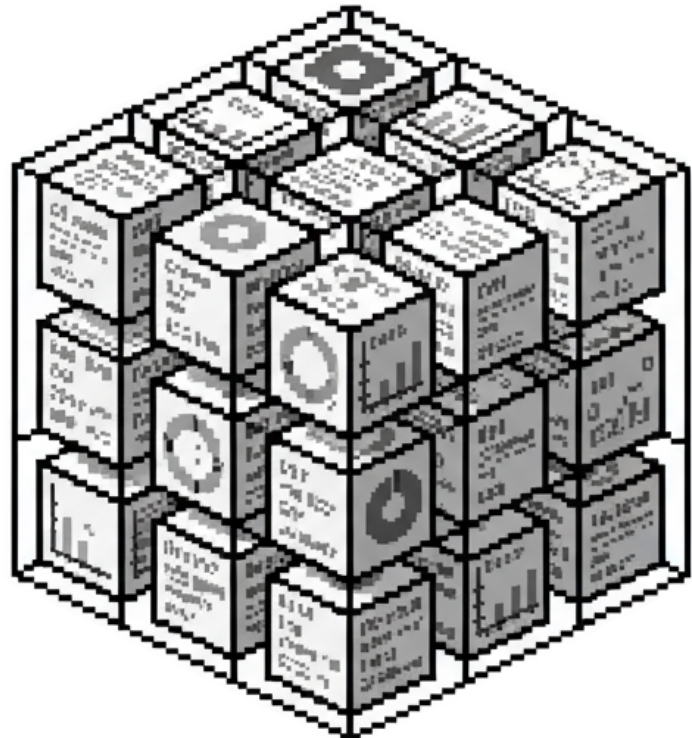
Мне нужен материал для катода аккумулятора с содержанием Никеля не более 70 процентов и удельной ёмкостью не менее 198 мАч/г



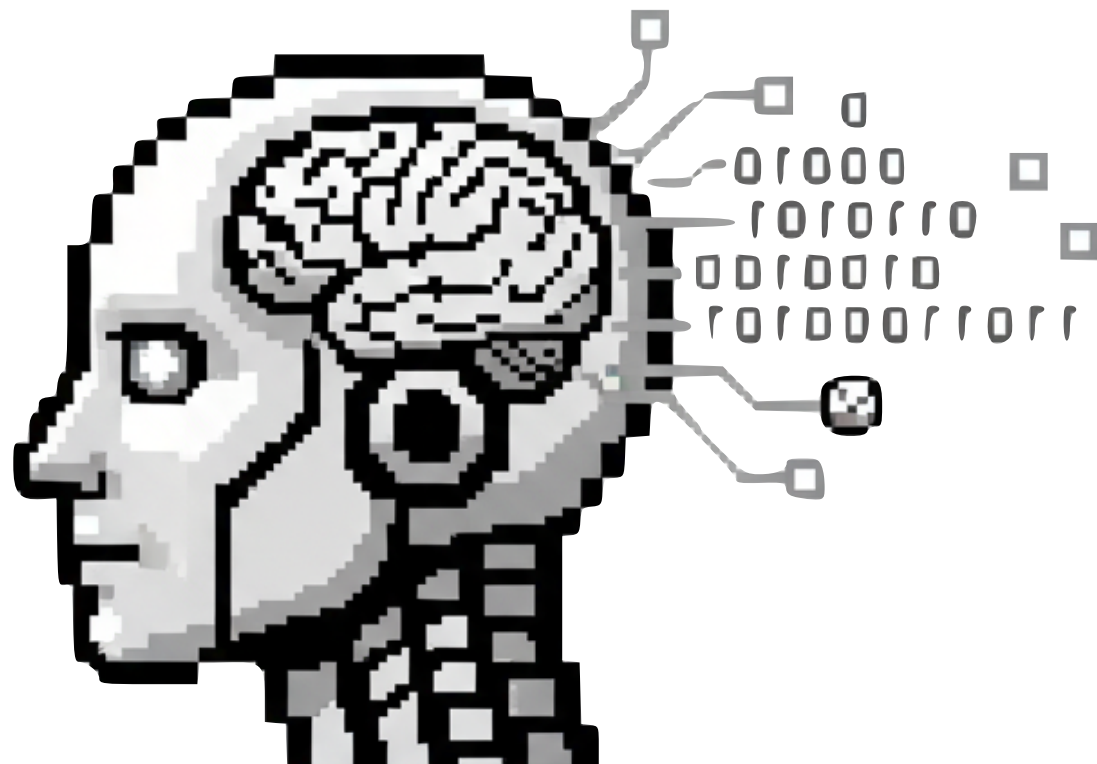
**База экспериментальных
данных**

Наличие такой базы экспериментальных данных может быть полезно для дизайна новых экспериментов

Мне нужен материал для катода аккумулятора с содержанием Никеля не более 70 процентов и удельной ёмкостью не менее 198 мАч/г



База экспериментальных данных

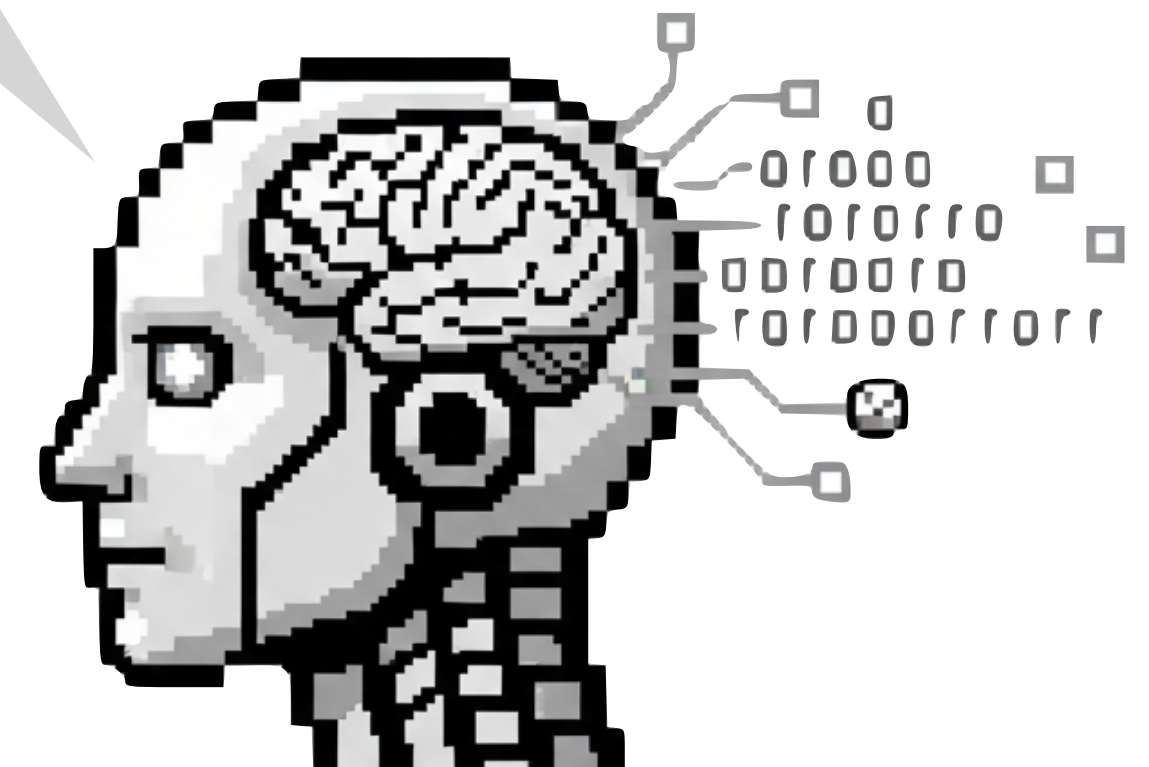


Математическая модель

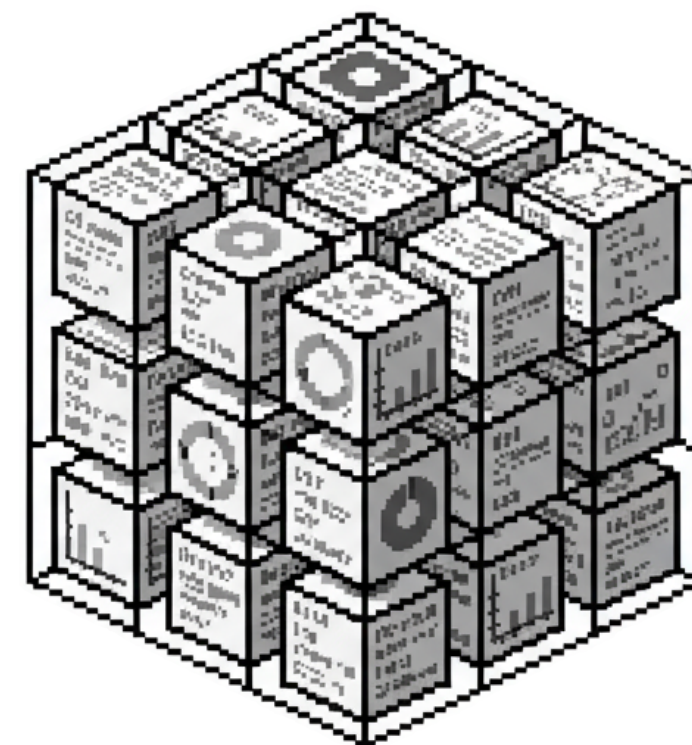
Наличие такой базы экспериментальных данных может быть полезно для дизайна новых экспериментов

Мне нужен материал для катода аккумулятора с содержанием Никеля не более 70 процентов и удельной ёмкостью не менее 198 мАч/г

Материалов с такими данными в базе нет. Судя по данным, имеет смысл поставить эксперимент с молекулой $\text{LiNi}_{0.5}\text{Mn}_{0.35}\text{Co}_{0.15}\text{O}_2$



Математическая модель



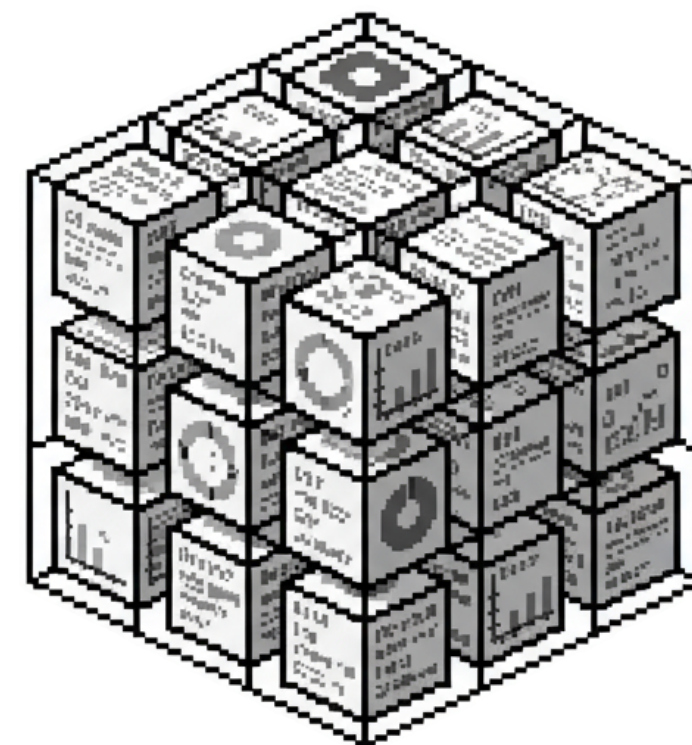
База экспериментальных данных



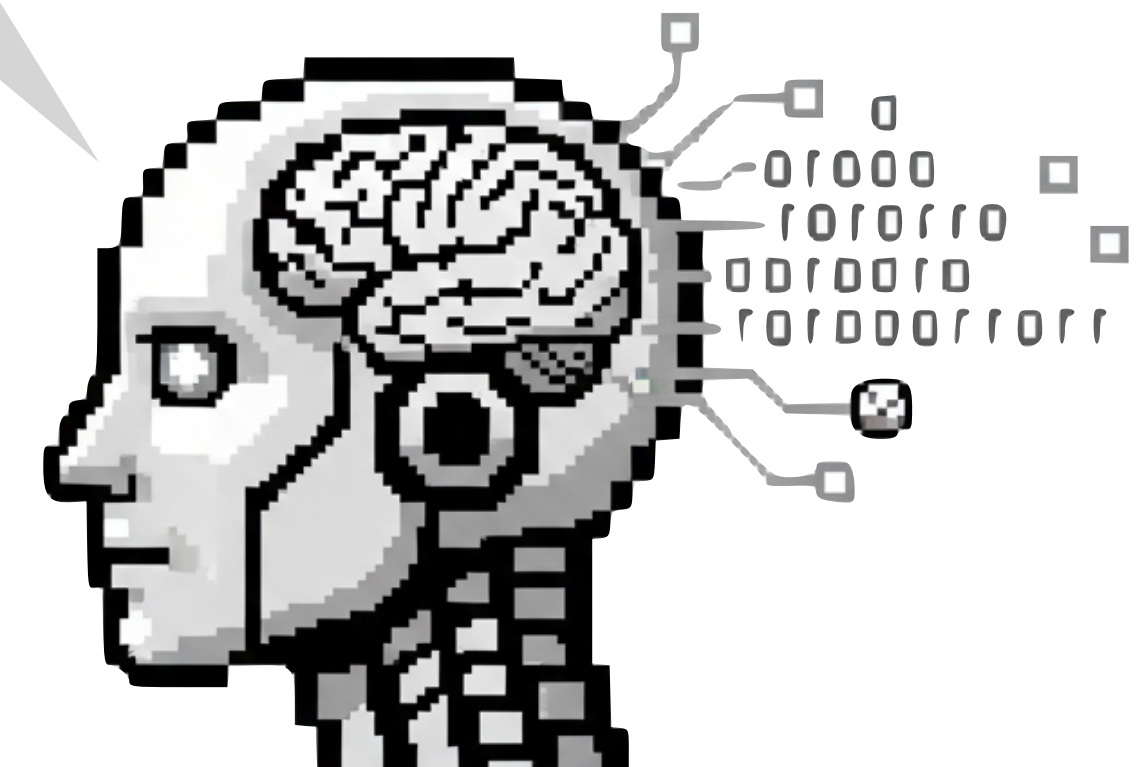
Наличие такой базы экспериментальных данных может быть полезно для дизайна новых экспериментов

Мне нужен материал для катода аккумулятора с содержанием Никеля не более 70 процентов и удельной ёмкостью не менее 198 мАч/г

Материалов с такими данными в базе нет. Судя по данным, имеет смысл поставить эксперимент с молекулой $\text{LiNi}_{0.5}\text{Mn}_{0.35}\text{Co}_{0.15}\text{O}_2$



База экспериментальных данных



Математическая модель

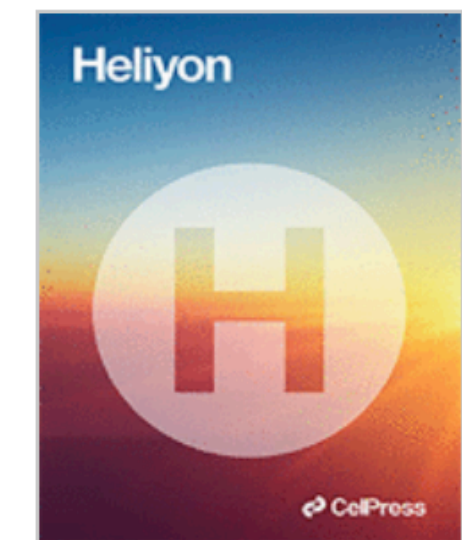
Проверка возможности на одном конкретном примере метаобзора

RESEARCH ARTICLE · Volume 9, Issue 12, e21881, December 2023 · *Open Access*

Benchmarking the electrochemical parameters of the $\text{LiNi}_{0.8}\text{Mn}_{0.1}\text{Co}_{0.1}\text{O}_2$ positive electrode material for Li-ion batteries

[Aleksandra A. Savina](#) · [Artem M. Abakumov](#)  

[Affiliations & Notes](#)  [Article Info](#) 



 [Download PDF](#)  [Cite](#)  [Set Alert](#)  [Get Rights](#)  [Reprints](#)

[< Previous article](#) [Next article >](#)

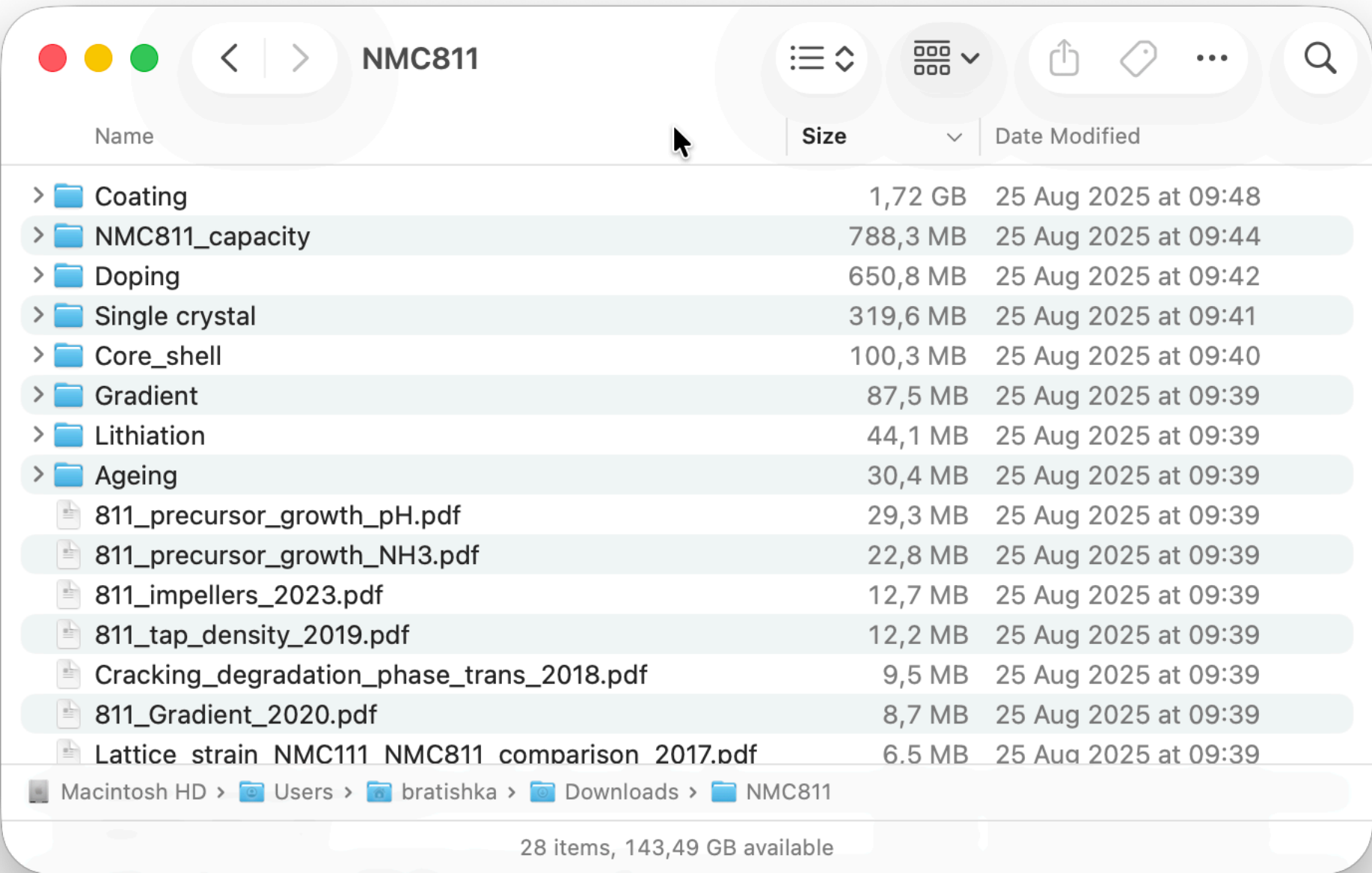
Попробуем воспроизвести автоматический парсинг научных статей и поиск релеватных данных научных экспериментов

Попытка автоматизировать процесс составления метаобзора области

Возьмём готовые
700+ статей у
авторов метаобзора

Распарсим .pdf/.doc
файлы в текст с
помощью docling/OCR

Ищем два главных
исследуемых параметра в
каждом тексте с помощью LLM



Только на скачивание всех статей ушло около одного часа.

3.89 Гбайт
743 статьи

Из скачанных статей надо выбрать только 500+ актуальных/
релевантных.

Этот шаг требует много работы для автоматизации

На первом этапе можно использовать научные поисковые
системы типа OpenAlex, SemanticScholar, однако, требует
дополнительной агентной помощи.

Планируем использовать модуль поиска релевантных
научных статей из платформы



Попытка автоматизировать процесс составления метаобзора области

Возьмём готовые
700+ статей у
авторов метаобзора

Распарсим .pdf/.doc
файлы в текст с
помощью docling/OCR

Ищем два главных
исследуемых параметра в
каждом тексте с помощью LLM



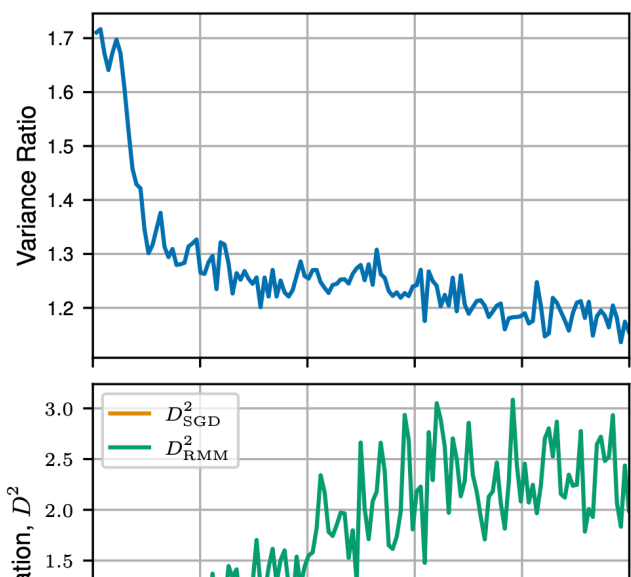
Memory-Efficient Backpropagation through Large Linear Layers

Table 2. Performance in fine-tuning on GLUE benchmark for different compression rates ρ (number of dimensions of projection space) in percents. The top row (No RMM) corresponds to a baseline implementation without compression.

	COLA	MNLI	MNLI-MM	MRPC	QNLI	QQP	RTE	SST2	STSB	WNLI	AVG
No RMM	60.90	87.56	87.24	88.24	92.62	91.69	78.34	94.95	90.68	56.34	82.86
90%	58.81	87.58	87.10	88.48	92.90	91.47	78.70	94.15	90.75	53.52	82.35
50%	58.60	87.58	87.25	88.97	92.93	91.41	79.42	94.15	90.85	56.34	82.75
20%	57.79	87.59	87.26	88.48	90.68	91.16	76.17	94.72	90.53	50.70	81.51
10%	56.52	87.51	87.28	87.01	92.40	90.93	71.84	94.38	90.29	43.66	80.18

Table 3. Maximal memory usage during training on GLUE tasks and memory economy for different compression rates ρ and a baseline implementation (No RMM).

TASK	BATCH	RATE	MEM, GiB	SAVING, %
MRPC	128	No RMM	11.3	0.0
		50%	10.6	6.3
		20%	9.2	19.3
		10%	8.7	23.3
QNLI	16	No RMM	11.7	0.0
		50%	11.2	4.2
		20%	10.4	11.6
		10%	10.1	13.8
SST2	256	No RMM	13.3	0.0
		50%	12.5	6.1
		20%	10.5	20.8
		10%	9.9	25.5



	COLA	MNLI	MNLI-MM	MRPC	QNLI	QQP	RTE	SST2	STSB	WNLI	AVG
No RMM	60.9	87.56	87.24	88.24	92.62	91.69	78.34	94.95	90.68	56.34	82.86
90%	58.81	87.58	87.1	88.48	92.9	91.47	78.7	94.15	90.75	53.52	82.35
50%	58.6	87.58	87.25	88.97	92.93	91.41	79.42	94.15	90.85	56.34	82.75
20%	57.79	87.59	87.26	88.48	90.68	91.16	76.17	94.72	90.53	50.7	81.51
10%	56.52	87.51	87.28	87.01	92.4	90.93	71.84	94.38	90.29	43.66	80.18

Table 3. Maximal memory usage during training on GLUE tasks and memory economy for different compression rates ρ and a baseline implementation (No RMM).

TASK	BATCH	RATE	MEM, GiB	SAVING, %
MRPC	128	No RMM	11.3	0
		50%	10.6	6.3
		20%	9.2	19.3
		10%	8.7	23.3
QNLI	16	No RMM	11.7	0
		50%	11.2	4.2
		20%	10.4	11.6
		10%	10.1	13.8
SST2	256	No RMM	13.3	0
		50%	12.5	6.1
		20%	10.5	20.8
		10%	9.9	25.5



Статьи преобразуются в текстовый файл с использованием разметки markdown.

Для работы используется docling. Поддерживаются разные входные форматы файлов: HTML, PDF, DOC и другие.

Работает парсинг таблиц. По умолчанию картинки не распознаются и не переводятся в текст. Однако, в научных вопросах и в составлении метаобзора это оказалось важным. Необходимо подключать детальную обработку графиков с помощью VLM и добавлять информацию из них как captioning в текст.

Парсинг всех статей в текст занимает ~2 часа

Попытка автоматизировать процесс составления метаобзора области

Возьмём готовые
700+ статей у
авторов метаобзора

Распарсим .pdf/.doc
файлы в текст с
помощью docling/OCR

Ищем два главных
исследуемых параметра в
каждом тексте с помощью LLM



```
Ты — эксперт по электрохимии, извлекающий числовые метрики из одной научной статьи в формате Markdown.
Требуется извлечь два показателя:
1) first-cycle discharge capacity — удельная разрядная ёмкость в ПЕРВОМ цикле (мА·ч/г, mAh/g, mAh g-1).
2) Coulombic efficiency (CE) — кулоновский КПД, предпочтительно в ПЕРВОМ цикле (initial/first-cycle CE, ICE), в
процентах.

Синонимы/варианты формулировок:
- first-cycle discharge capacity: "first(-)cycle discharge specific capacity", "initial discharge capacity",
"first discharge capacity", "first cycle specific capacity (discharge)", "initial specific discharge capacity".
- Единицы: "mAh/g", "mAh g-1", "mA h g-1", "mA·h g-1".
- CE: "Coulombic efficiency", "first-cycle CE", "initial Coulombic efficiency", "ICE", "initial CE".

Правила:
- Ищи как в тексте, так и в таблицах/подписях к рисункам.
- Для discharge capacity брать ИМЕННО разрядную ёмкость первого цикла. Если указана только "first cycle capacity"
в таблице Charge/Discharge, возьми значение из колонки Discharge. Игнорируй charge capacity, average/after N
cycles и reversible capacity, если явно не сказано, что это first-cycle DISCHARGE.
- Для CE приоритет: initial/first-cycle CE. Если есть диапазон или несколько значений, возьми первое явно
подписанное как initial/first-cycle. Если дано в доле (0.945), верни в процентах (94.5).
- Возвращай только числа без единиц: capacity — в mAh/g; CE — в процентах (0-100, можно с десятичной точкой).
- Если показатель не найден надёжно, верни null.
- Заголовок статьи возьми из первого h1 в Markdown, либо используй переданный title_guess, если h1 отсутствует.

Формат ответа: один JSON-объект БЕЗ комментариев:
{
  "title": string,
  "first_cycle_discharge_capacity_mAh_g": number | null,
  "coulombic_efficiency_percent": number | null
}
Никакого дополнительного текста.
| %L to chat, %K to generate
```

```
Найдено файлов: 751
0%|          | 0/751 [00:00<?, ?it/s]
[1/751] OK: 811_impellers_2023.md -> FCDC=100.12 mAh/g, CE=96.02 %
0%|          | 1/751 [00:09<1:58:10, 9.45s/it]
[2/751] OK: NMC811_EELS.md -> FCDC=None mAh/g, CE=None %
0%|          | 2/751 [00:27<3:02:09, 14.59s/it]
[3/751] OK: Antisite_NMC811_2020.md -> FCDC=187.0 mAh/g, CE=None %
0%|          | 3/751 [00:52<4:01:57, 19.41s/it]
[4/751] OK: 811_tap_density_2019.md -> FCDC=203.7 mAh/g, CE=86.3 %
1%|          | 4/751 [01:16<4:20:36, 20.93s/it]
```

Запуск локальной LLM GPT-oss 120b для
извлечения важных параметров из текста на Nvidia
A100

Обработка всех статей занимает около 4 часов

Пока никак не учитывается содержимое картинок,
что немного мешает проверке



Попытка автоматизировать процесс составления метаобзора области

Возьмём готовые
700+ статей у
авторов метаобзора

Распарсим .pdf/.doc
файлы в текст с
помощью docling/OCR

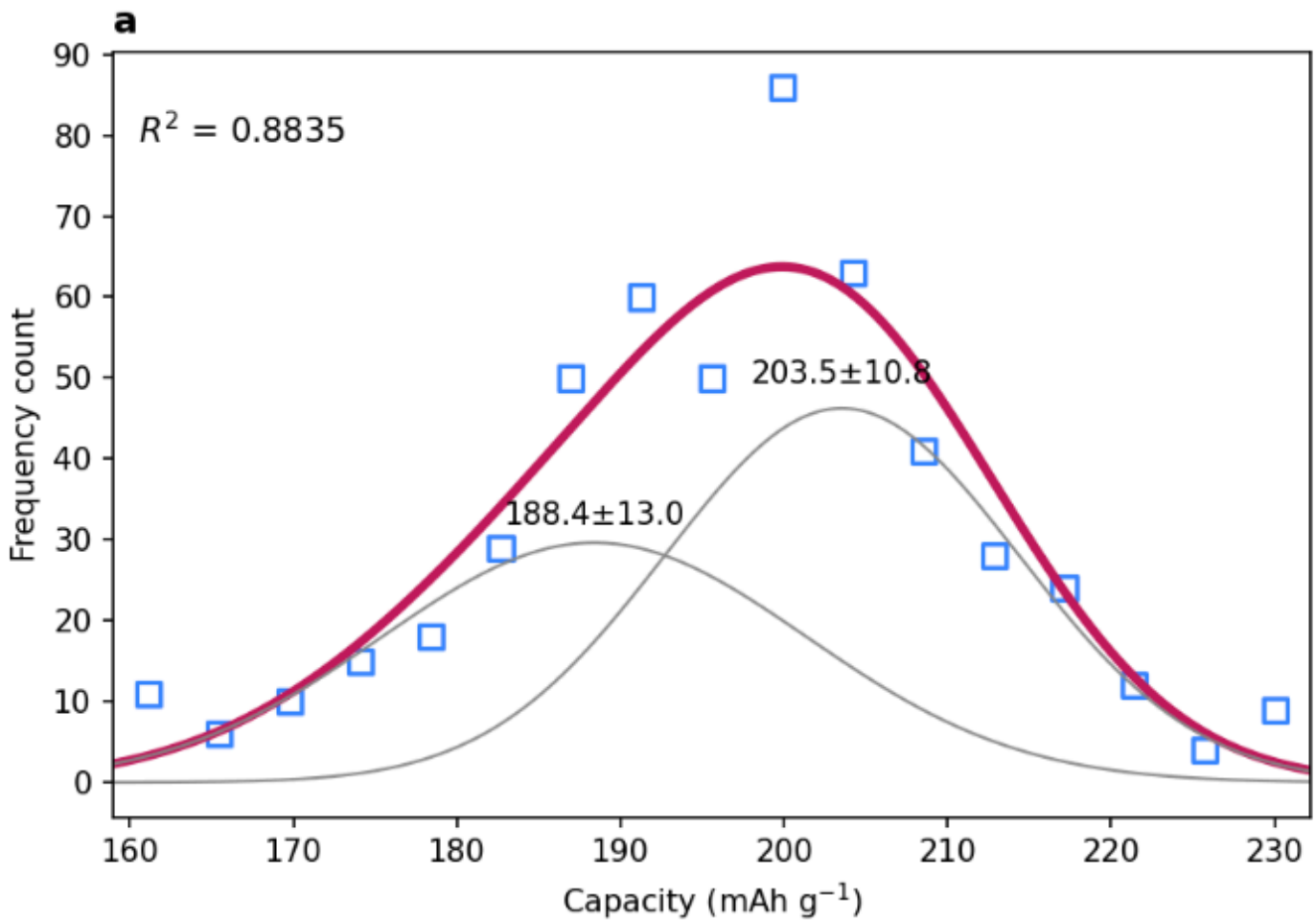
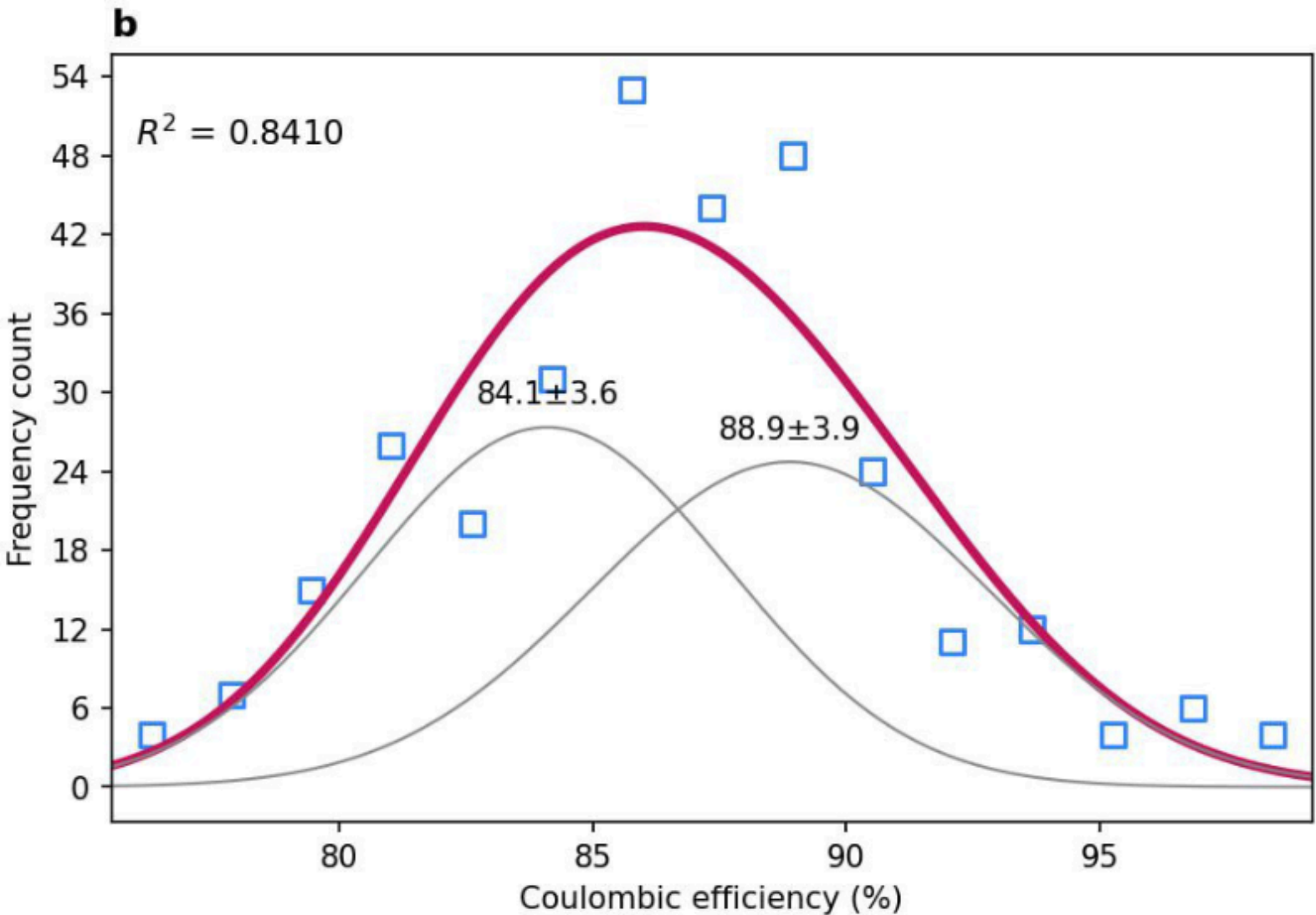
Ищем два главных
исследуемых параметра в
каждом тексте с помощью LLM

1 час

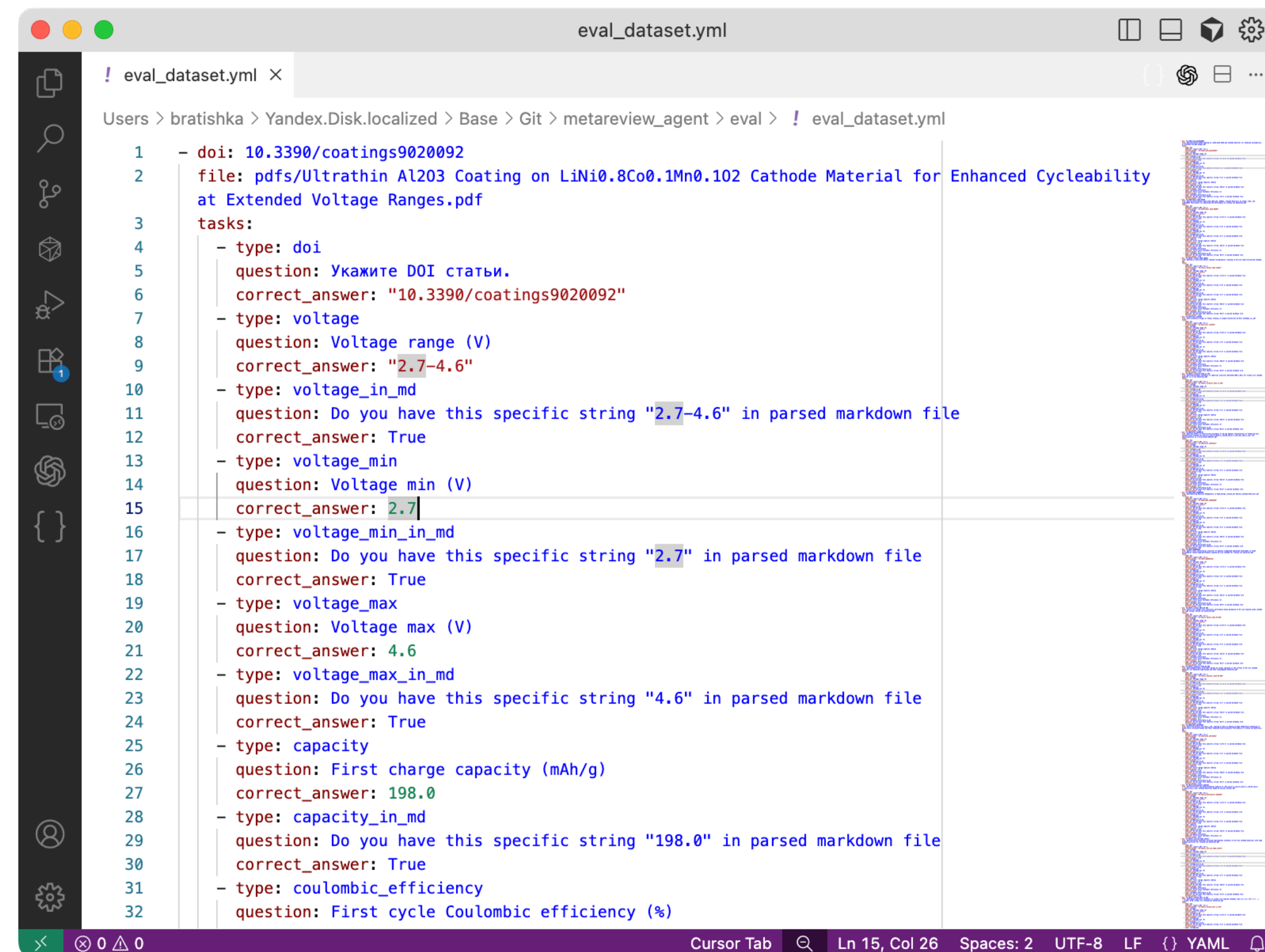
2 час

4 часа

7 часов работы



Оценка качества полученного решения



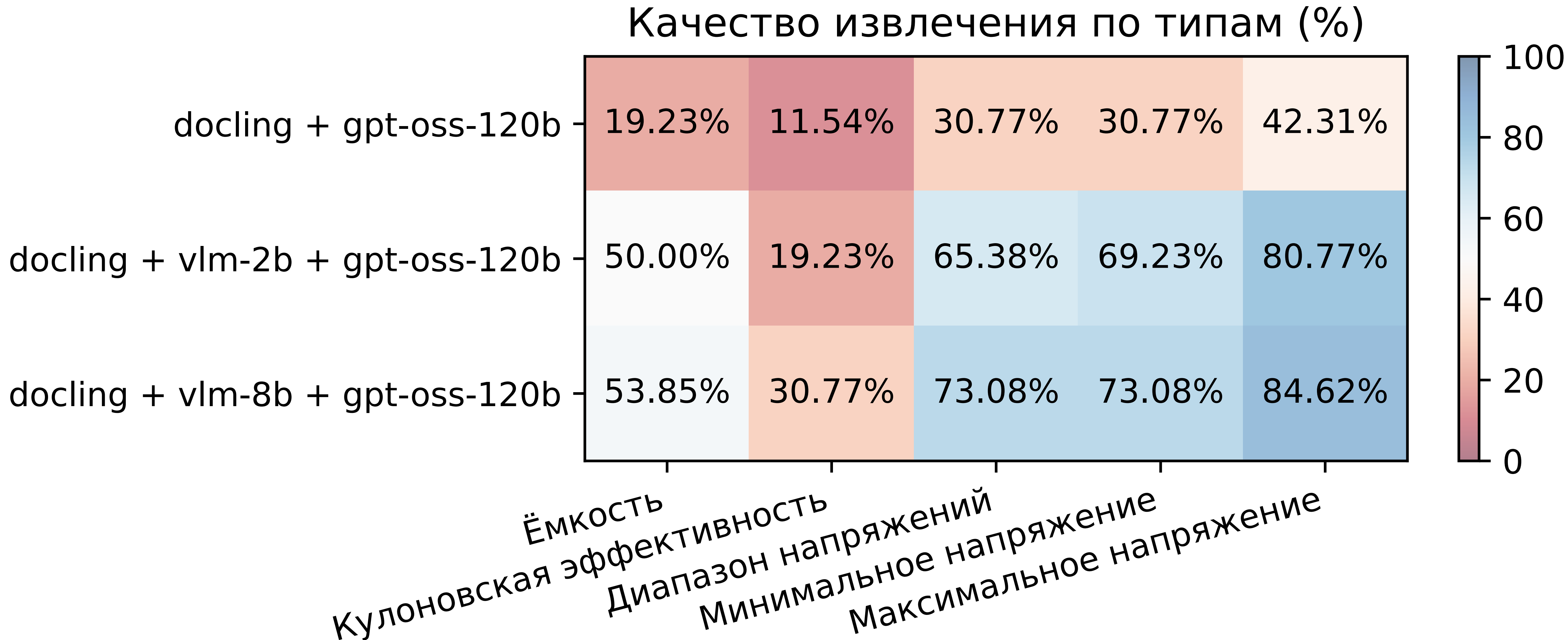
```
1 - doi: 10.3390/coatings9020092
2   file: pdfs/Ultrathin Al2O3 Coating on LiNi0.8Co0.1Mn0.102 Cathode Material for Enhanced Cycleability
3   at Extended Voltage Ranges.pdf
4   tasks:
5     - type: doi
6       question: Укажите DOI статьи.
7       correct_answer: "10.3390/coatings9020092"
8     - type: voltage
9       question: Voltage range (V)
10      correct_answer: "2.7-4.6"
11     - type: voltage_in_md
12       question: Do you have this specific string "2.7-4.6" in parsed markdown file
13       correct_answer: True
14     - type: voltage_min
15       question: Voltage min (V)
16       correct_answer: 2.7
17     - type: voltage_min_in_md
18       question: Do you have this specific string "2.7" in parsed markdown file
19       correct_answer: True
20     - type: voltage_max
21       question: Voltage max (V)
22       correct_answer: 4.6
23     - type: voltage_max_in_md
24       question: Do you have this specific string "4.6" in parsed markdown file
25       correct_answer: True
26     - type: capacity
27       question: First charge capacity (mAh/g)
28       correct_answer: 198.0
29     - type: capacity_in_md
30       question: Do you have this specific string "198.0" in parsed markdown file
31       correct_answer: True
32     - type: coulombic_efficiency
33       question: First cycle Coulombic efficiency (%)
```

Собрали датасет для измерения качества извлечения релевантной научной информации из скачанных статей из 26 статей

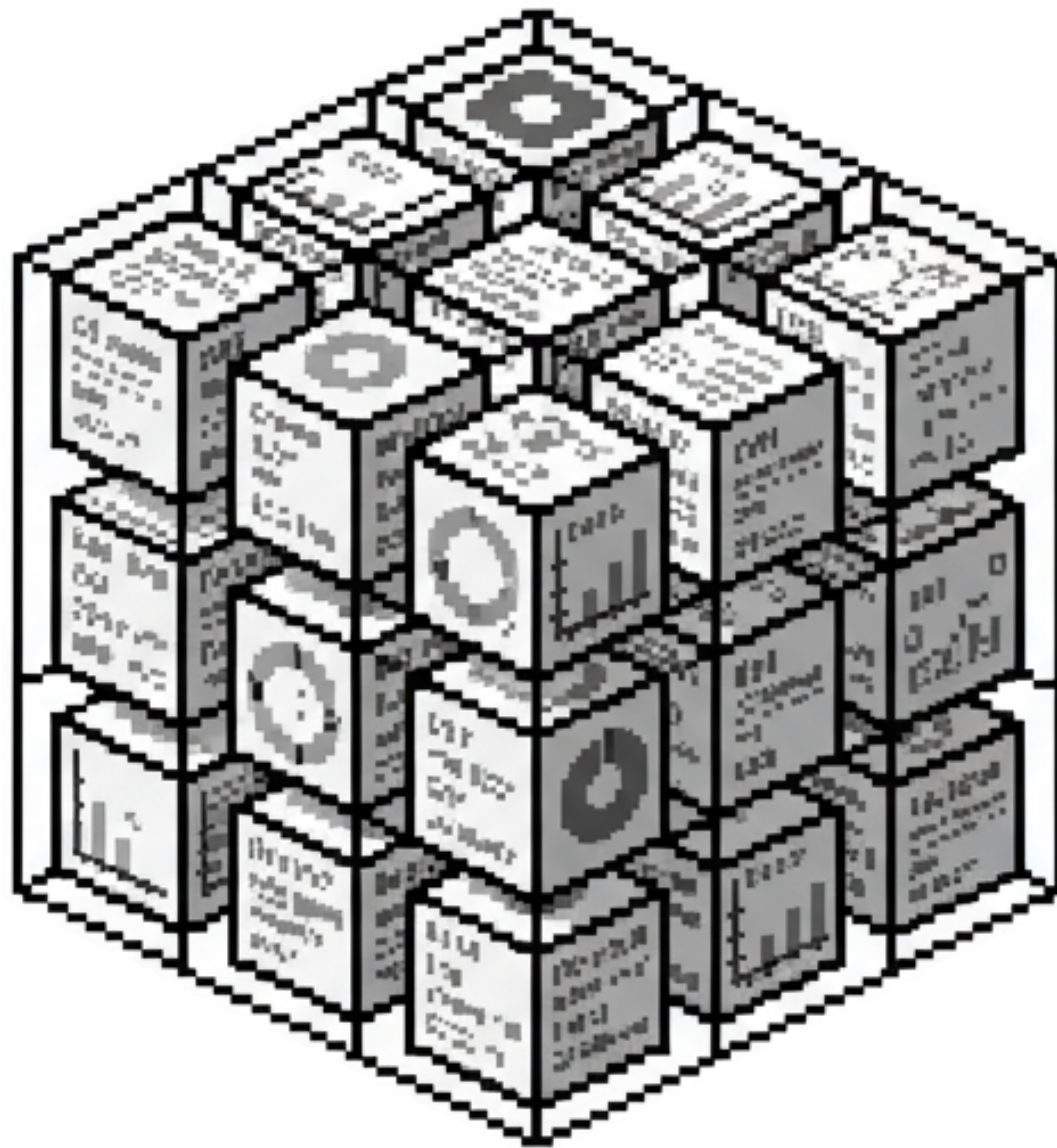
Наивный подход к задаче не обеспечивает приемлемое качество

-  Поиск релевантных статей
-  Парсинг текста
-  Парсинг таблиц
-  Поиск релевантных параметров из распознанного текста
-  Поиск релевантных параметров из графиков

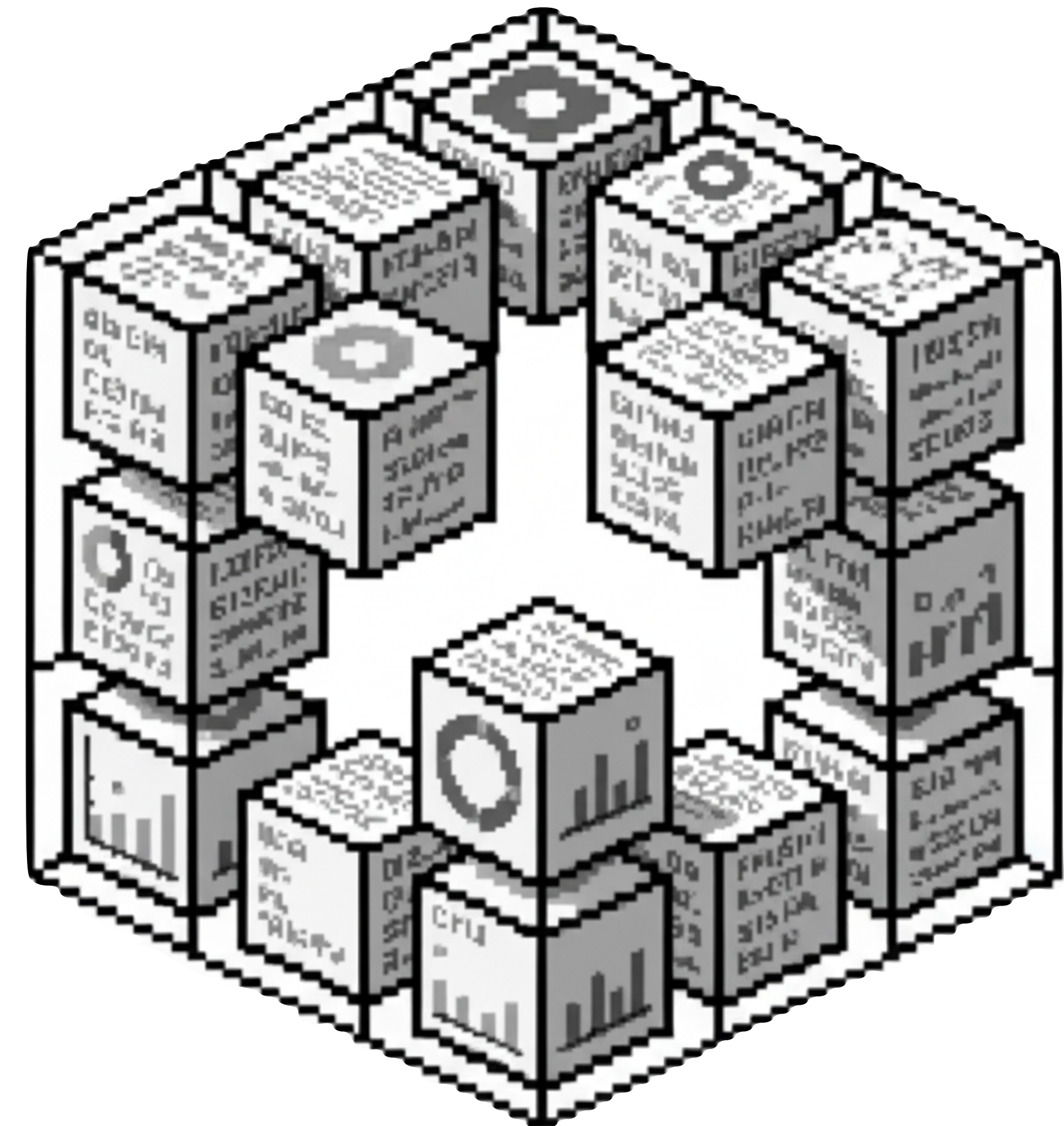
Качество извлечения информации на собранном тестовом датасете



Извлечение знания из неполных экспериментальных данных

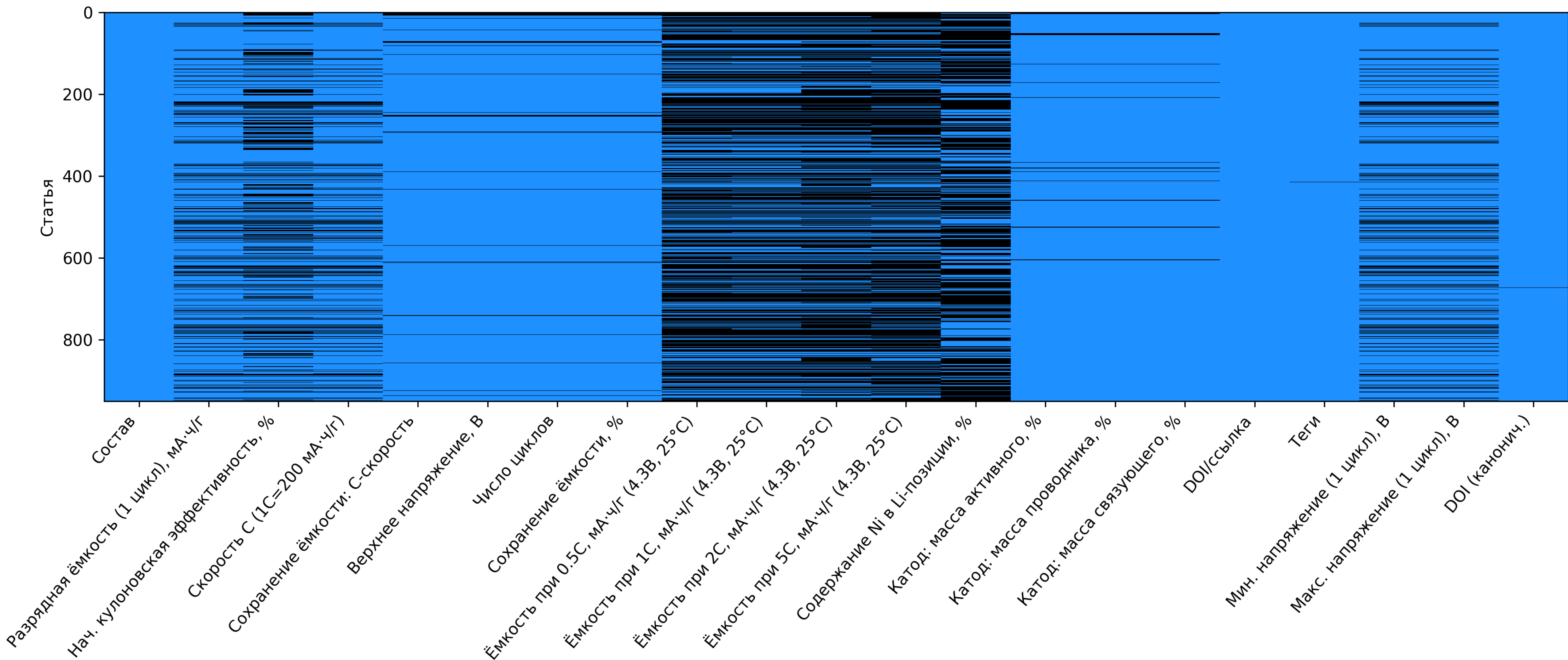


Так не бывает почти никогда

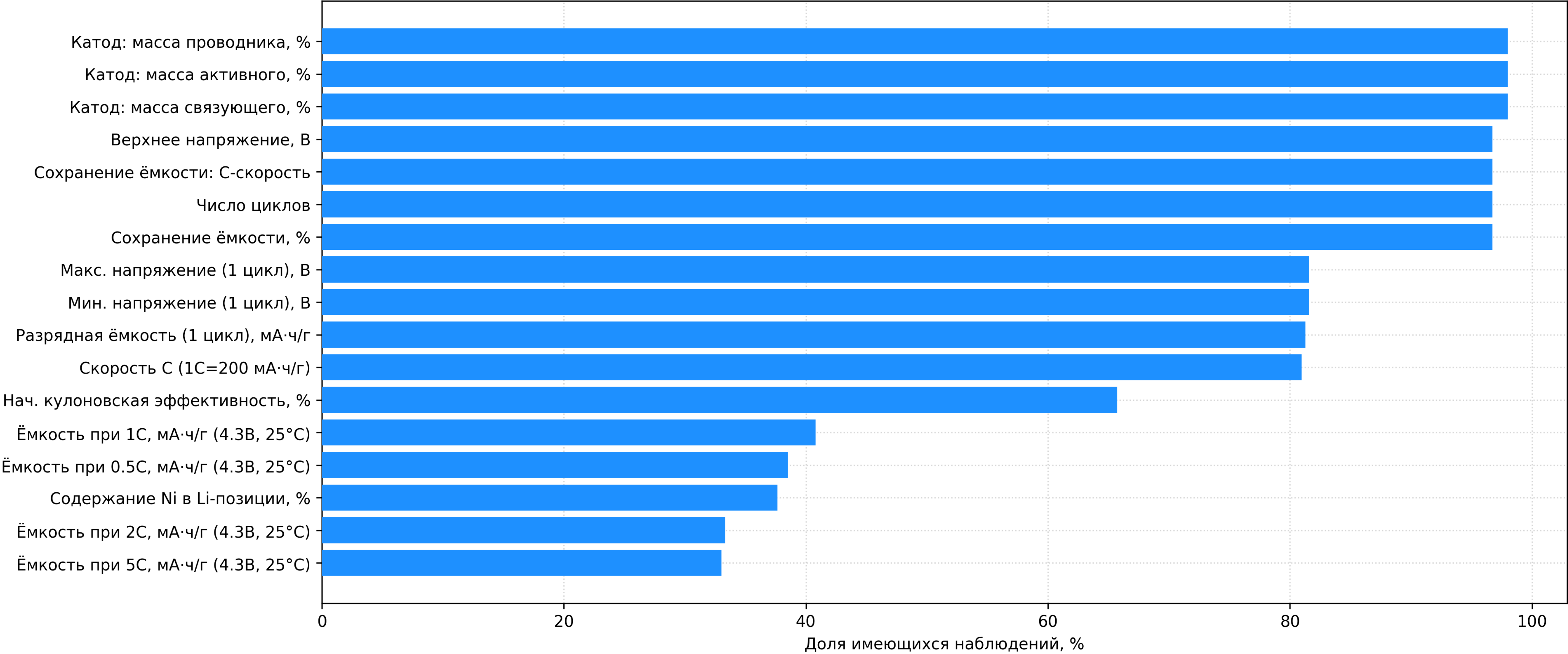


Обычно в базе данных есть пропуски. Имея часть данных, можно принимать более информированное решение о дизайне новых экспериментов

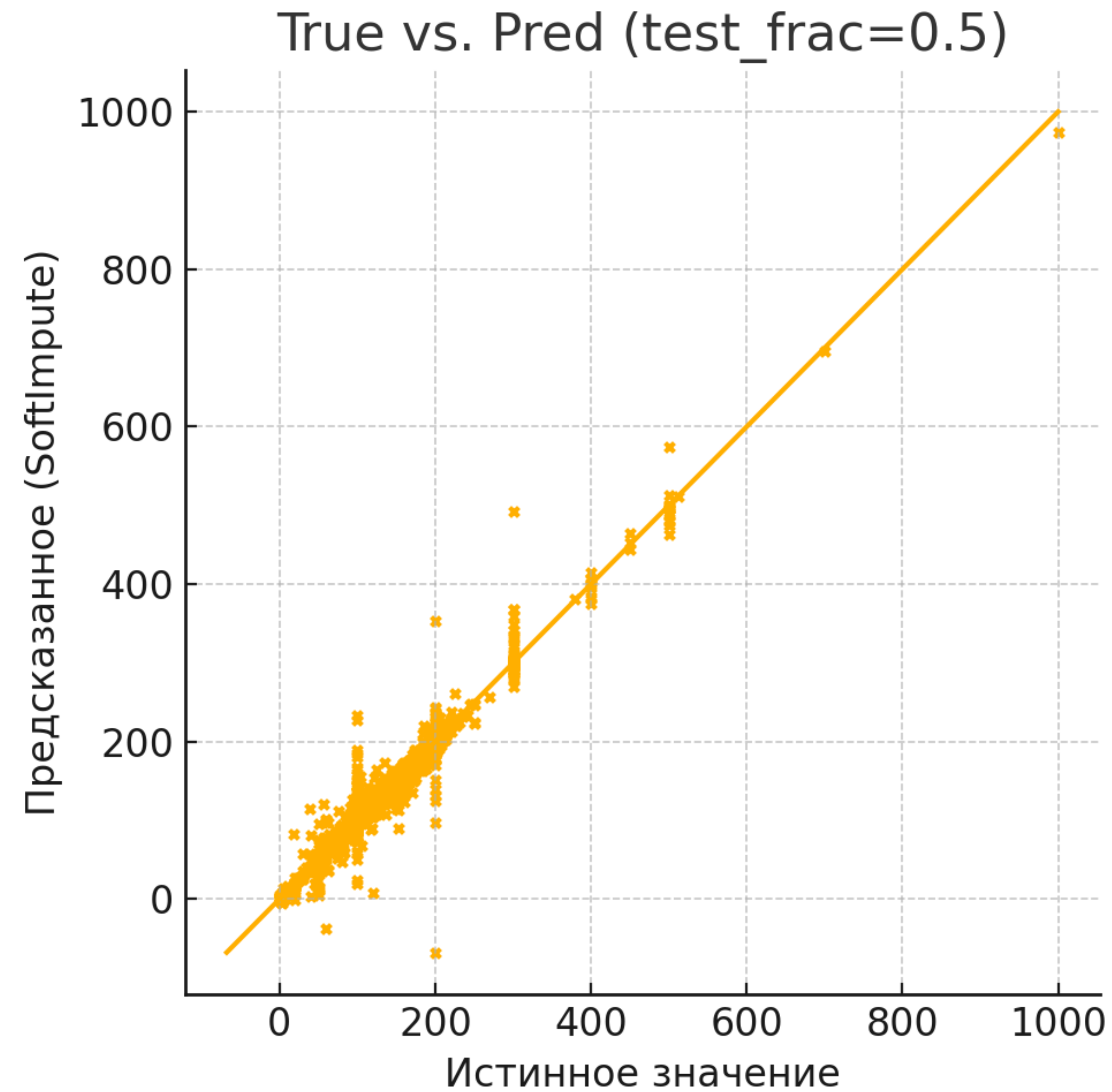
Даже в оригинальном датасете много пропущенных значений (черный цвет)



Даже в оригинальном датасете много пропущенных значений (черный цвет)



Подход matrix completion



Текущие слабые стороны подхода и точки роста

- Прямо сейчас мы собираем (или берём готовый) набор статей в области руками
 - *Потенциальное решение:* использование модуля поиска релевантных научных статей из платформы + систем научного индексирования (внутренних или внешних)
- Для получения информации о том, какие именно параметры необходима экспертиза в доменной области
 - *Потенциальное решение:* построение агента, который может в диалоговом режиме уточнения анализировать метаинформацию работ в области и формулировать наиболее важные параметры для экстракции
- Возникает необходимость детального анализа графиков в статьях с тем, чтобы вычленять из них экстремумы, наклоны, соотношения, другую метаинформацию - задача действительно непростая. Однако, часто в статьях ключевые выводы дублируются текстом.
 - *Потенциальное решение:* использовать VLM для анализа графиков и другие подходы Chart Question Answering

Развитие подхода



Схема работы научного поиска в области с экспериментами

Мне нужен материал для катода аккумулятора с содержанием Никеля не более 70 процентов и удельной ёмкостью не менее 198 мАч/г



Схема работы научного поиска в области с экспериментами

Мне нужен материал для катода аккумулятора с содержанием Никеля не более 70 процентов и удельной ёмкостью не менее 198 мАч/г



Исследование (от недели до нескольких месяцев)

Схема работы научного поиска в области с экспериментами

Мне нужен материал для катода аккумулятора с содержанием Никеля не более 70 процентов и удельной ёмкостью не менее 198 мАч/г

Поиск статей



Исследование (от недели до нескольких месяцев)

Схема работы научного поиска в области с экспериментами

Мне нужен материал для катода аккумулятора с содержанием Никеля не более 70 процентов и удельной ёмкостью не менее 198 мАч/г



Поиск статей

Фильтрация результатов



Исследование (от недели до нескольких месяцев)

Схема работы научного поиска в области с экспериментами

Мне нужен материал для катода аккумулятора с содержанием Никеля не более 70 процентов и удельной ёмкостью не менее 198 мАч/г



Поиск статей

Выбор наиболее подходящего материала из имеющегося

Фильтрация результатов



Исследование (от недели до нескольких месяцев)

Схема работы научного поиска в области с экспериментами

Мне нужен материал для катода аккумулятора с содержанием Никеля не более 70 процентов и удельной ёмкостью не менее 198 мАч/г



Поиск статей

Поиск
метаобзоров

Фильтрация
результатов

Выбор наиболее подходящего материала из имеющегося



Исследование (от недели до нескольких месяцев)

Схема работы научного поиска в области с экспериментами

Мне нужен материал для катода аккумулятора с содержанием Никеля не более 70 процентов и удельной ёмкостью не менее 198 мАч/г



Поиск статей Формулировка требований Выбор наиболее подходящего материала из имеющегося

Поиск метаобзоров Фильтрация результатов



Исследование (от недели до нескольких месяцев)

Схема работы научного поиска в области с экспериментами

Мне нужен материал для катода аккумулятора с содержанием Никеля не более 70 процентов и удельной ёмкостью не менее 198 мАч/г



Поиск статей Формулировка требований Выбор наиболее подходящего материала из имеющегося

Поиск метаобзоров Фильтрация результатов Оценка качества работ



Исследование (от недели до нескольких месяцев)

Схема работы научного поиска в области с экспериментами

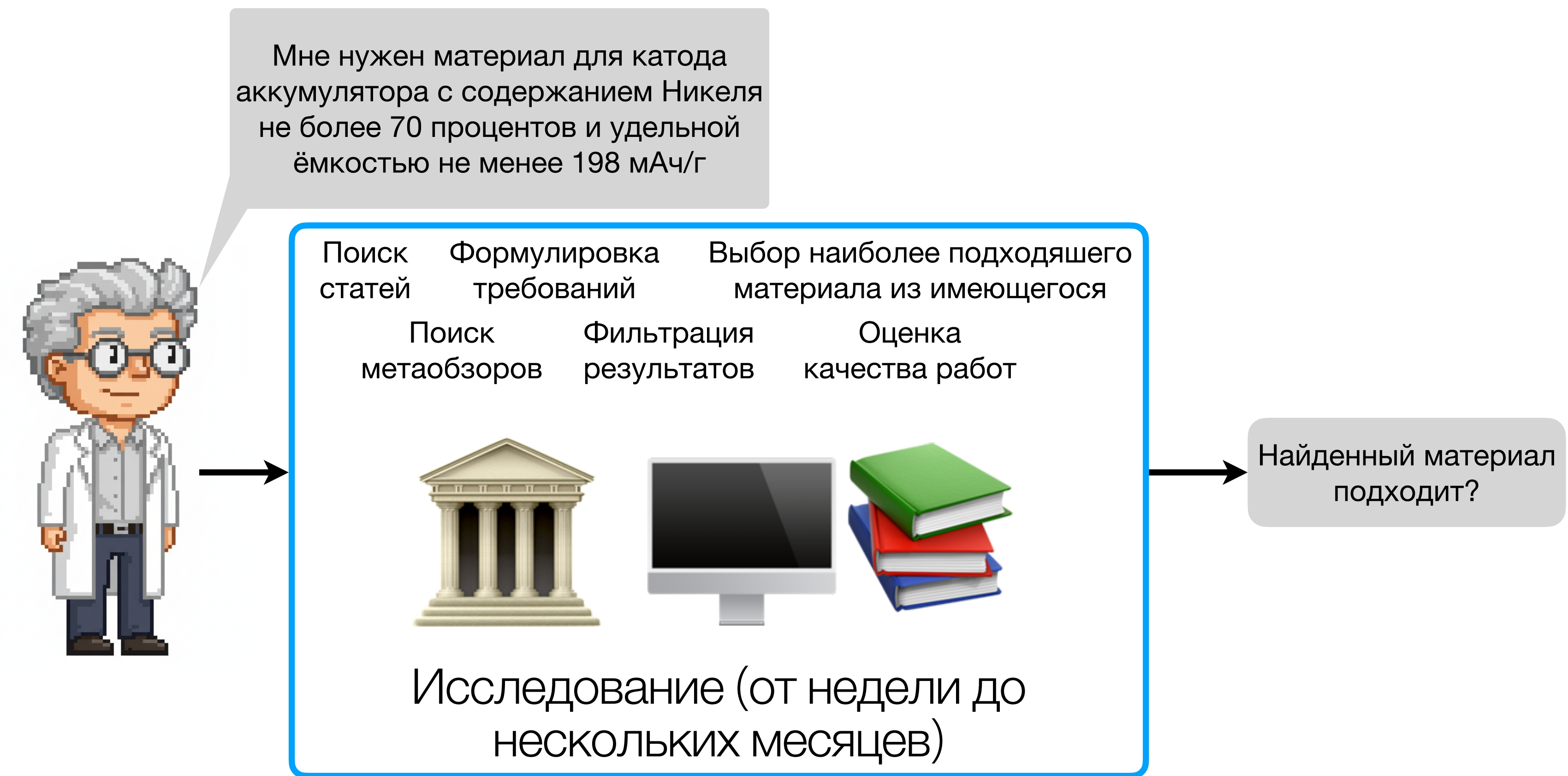


Схема работы научного поиска в области с экспериментами

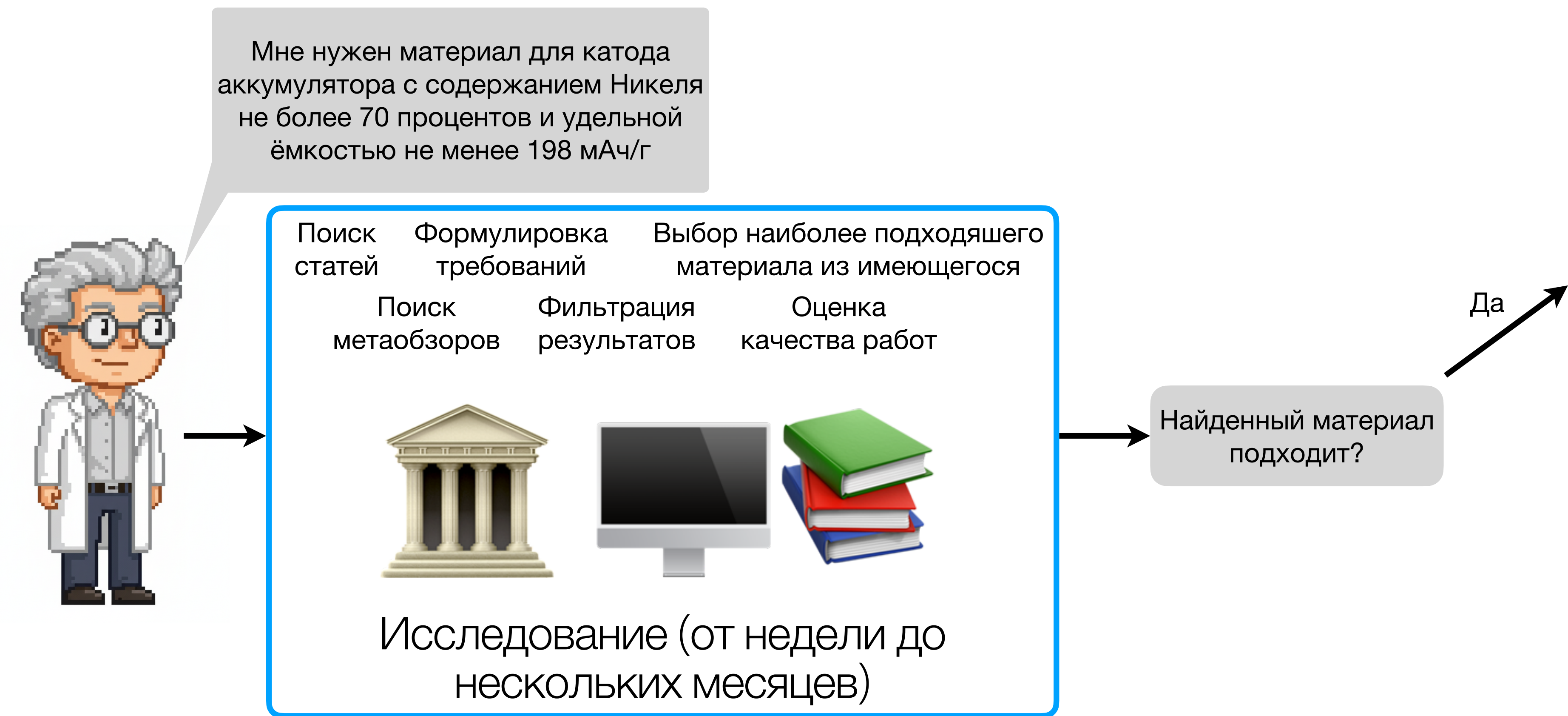


Схема работы научного поиска в области с экспериментами

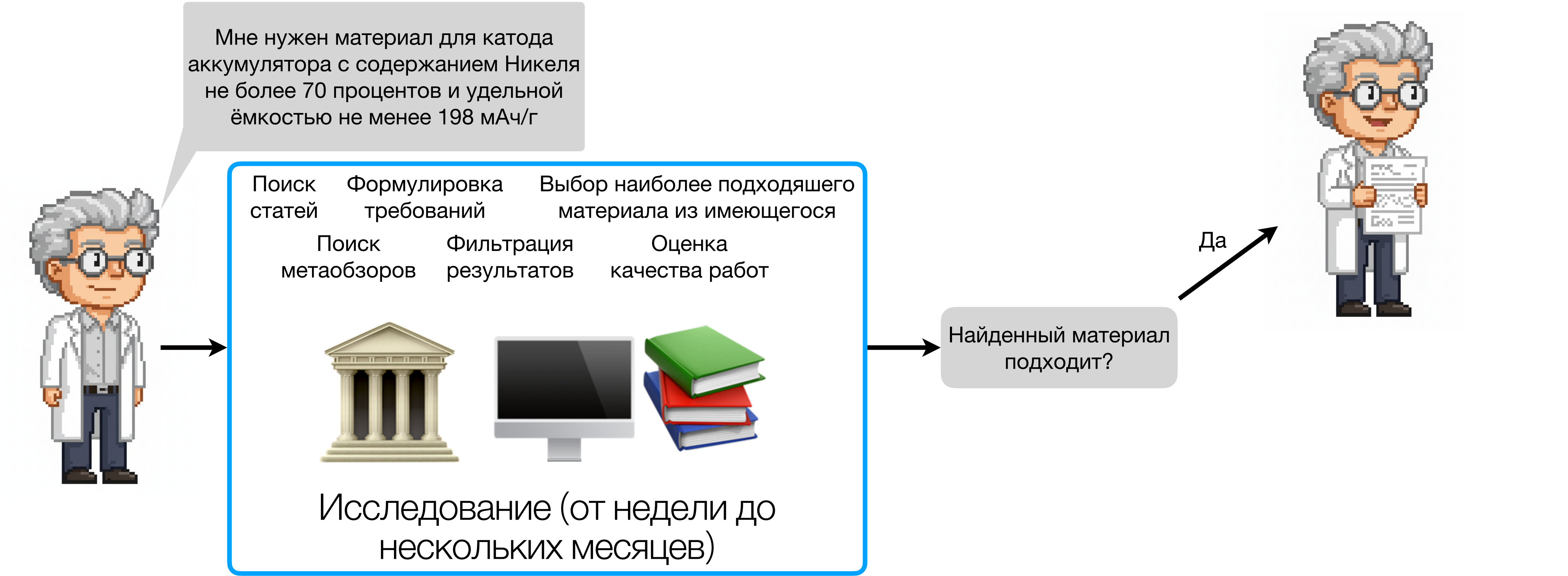


Схема работы научного поиска в области с экспериментами

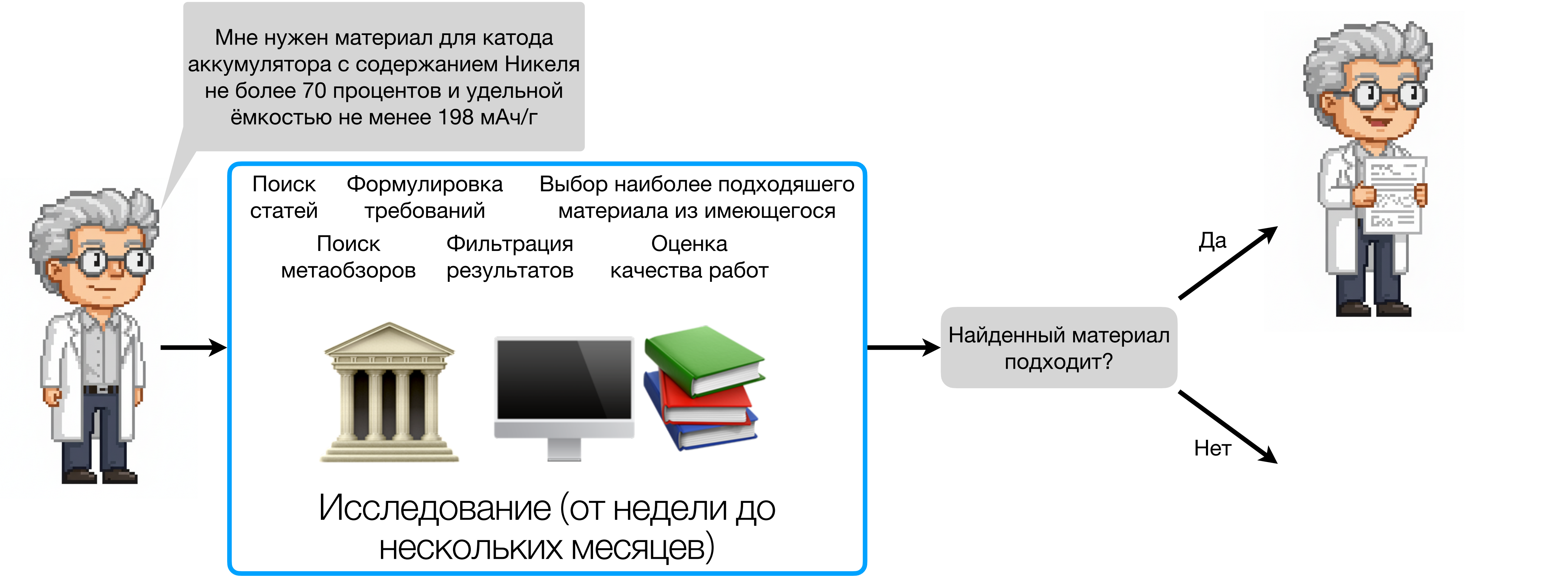


Схема работы научного поиска в области с экспериментами

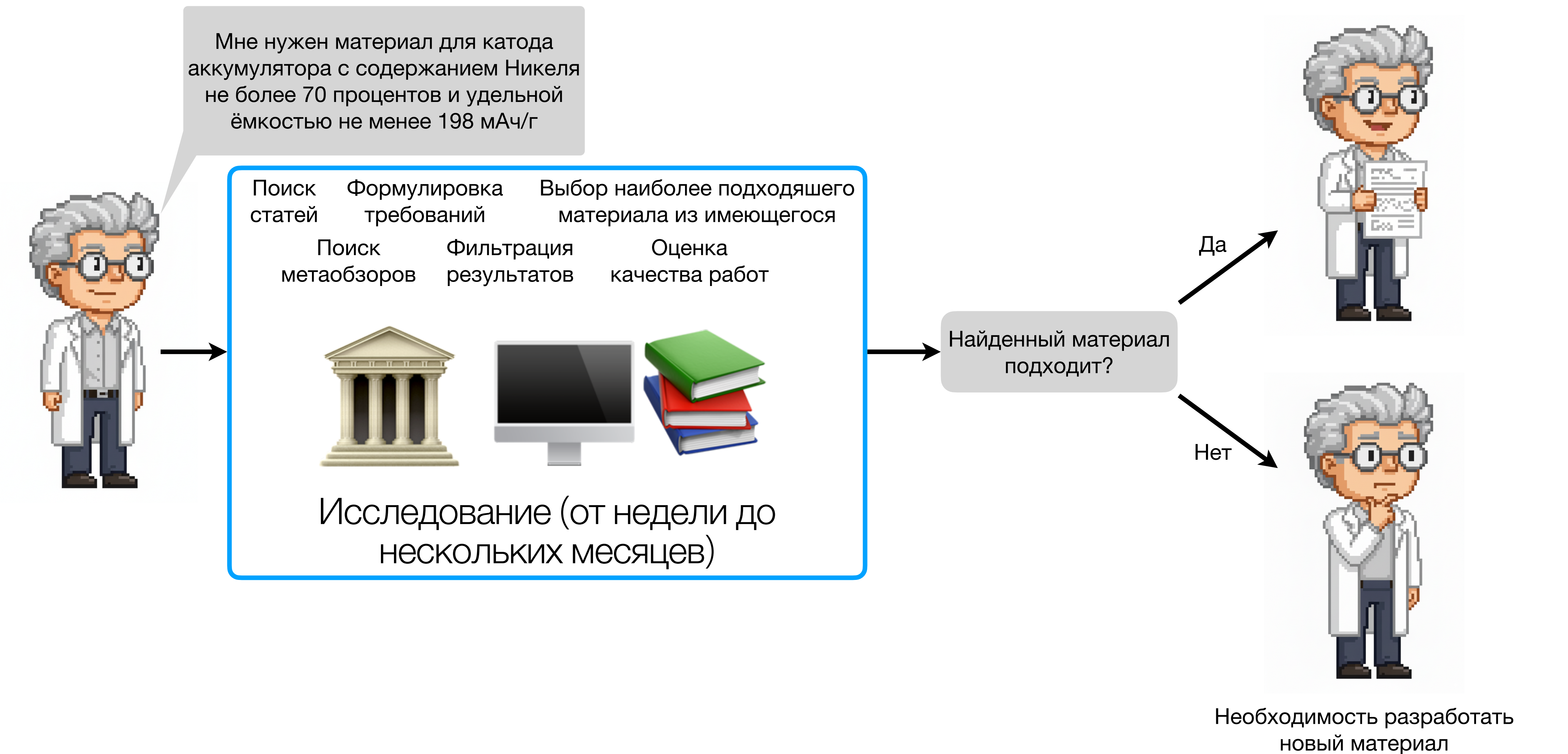


Схема работы научного поиска в области с экспериментами

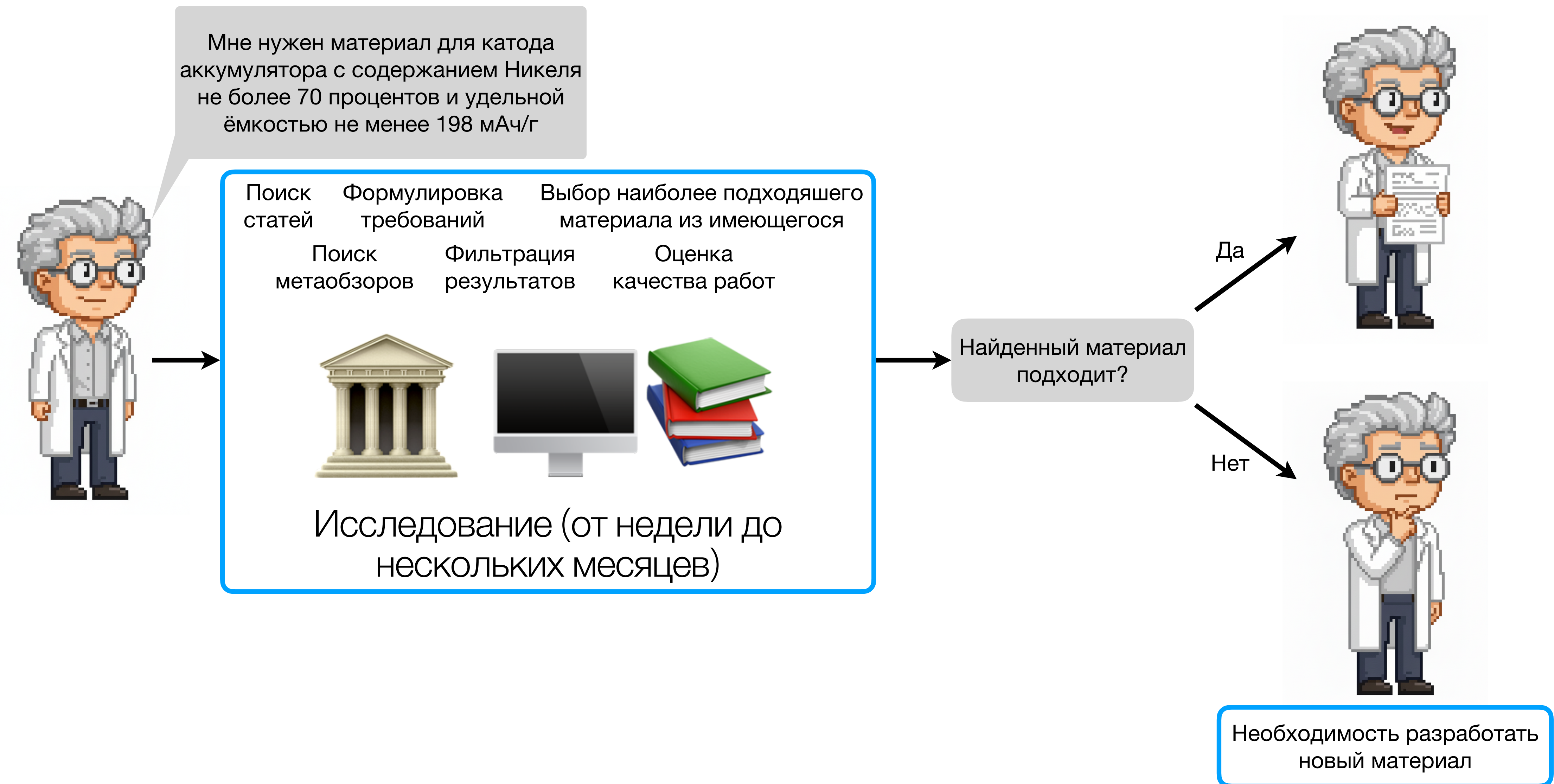
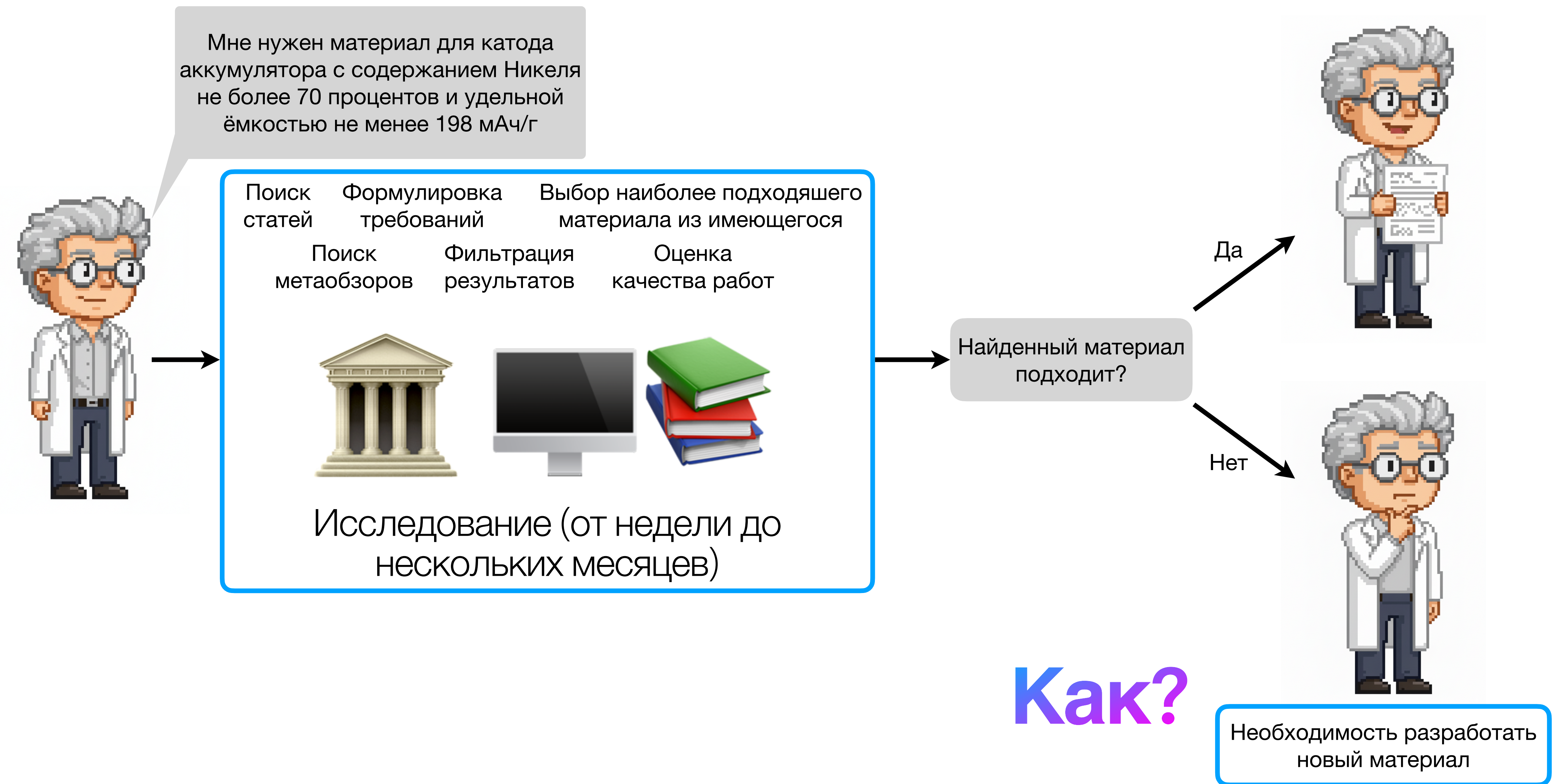
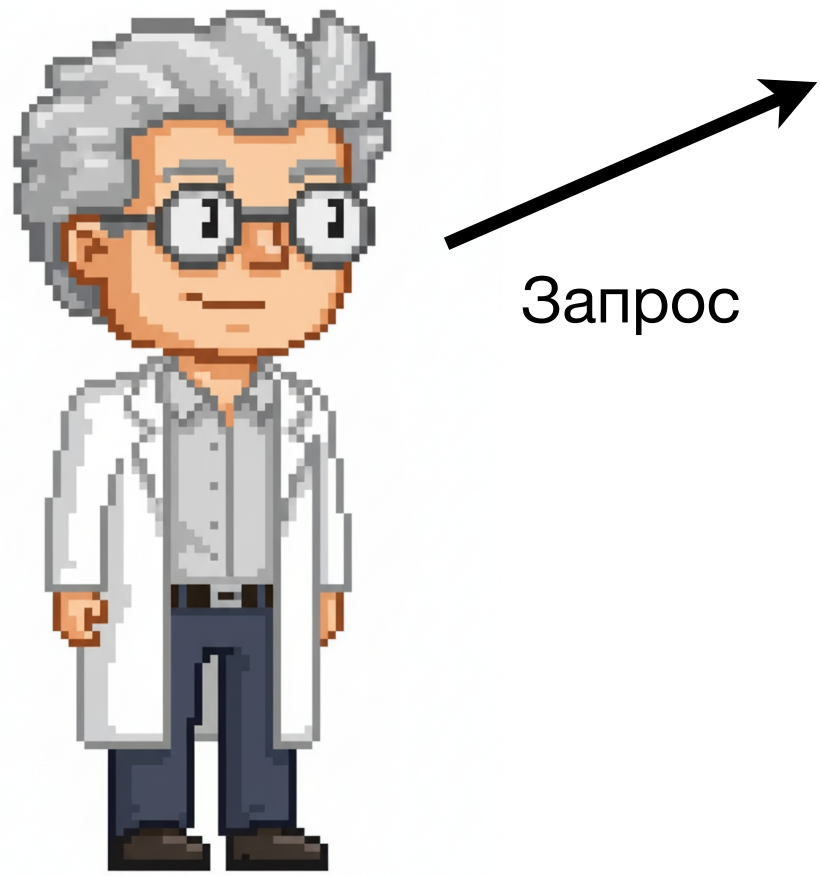


Схема работы научного поиска в области с экспериментами



Data-driven science



Data-driven science



Запрос

Агент формулирует
наиболее важные/
релевантные параметры
для исследования
заданной области

Например, для конкретного обзора мы потратили около часа разговора с экспертом для того, чтобы понять какие параметры из каждой статьи нам лучше всего экстрактить с помощью LLM.

Текущие системы Deep Research уже неплохо справляются с этой задачей.

Полученный агент может быть полезен для разных областей знания

Data-driven science



Запрос

Агент формулирует
наиболее важные/
релевантные параметры
для исследования
заданной области

Поиск релевантных
научных статей

Data-driven science



Запрос

Агент формулирует
наиболее важные/
релевантные параметры
для исследования
заданной области

Поиск релевантных
научных статей

Поиск нужных
параметров в скачанных
научных статьях

Data-driven science



Запрос

Агент формулирует
наиболее важные/
релевантные параметры
для исследования
заданной области

Поиск релевантных
научных статей

Поиск нужных
параметров в скачанных
научных статьях

Составление/
обновление базы
данных в области

Data-driven science



Запрос

Агент формулирует наиболее важные/ релевантные параметры для исследования заданной области

Поиск релевантных научных статей

Поиск нужных параметров в скачанных научных статьях

Построение математической модели собранных данных

Составление/ обновление базы данных в области

Собранный датасет в конечном итоге представляет из себя многомерную массив данных (тензор). Мы можем исследовать базовые подходы tensor completion для того, чтобы подсказывать исследователю недостающие части этой таблицы, что приводит к data-driven дизайну экспериментов.

Кроме того, полученные данные в результате могут помочь перебирать меньшее количество вариантов для экспериментальных исследований.

Data-driven science



Запрос

Агент формулирует наиболее важные/ релевантные параметры для исследования заданной области

Поиск релевантных научных статей

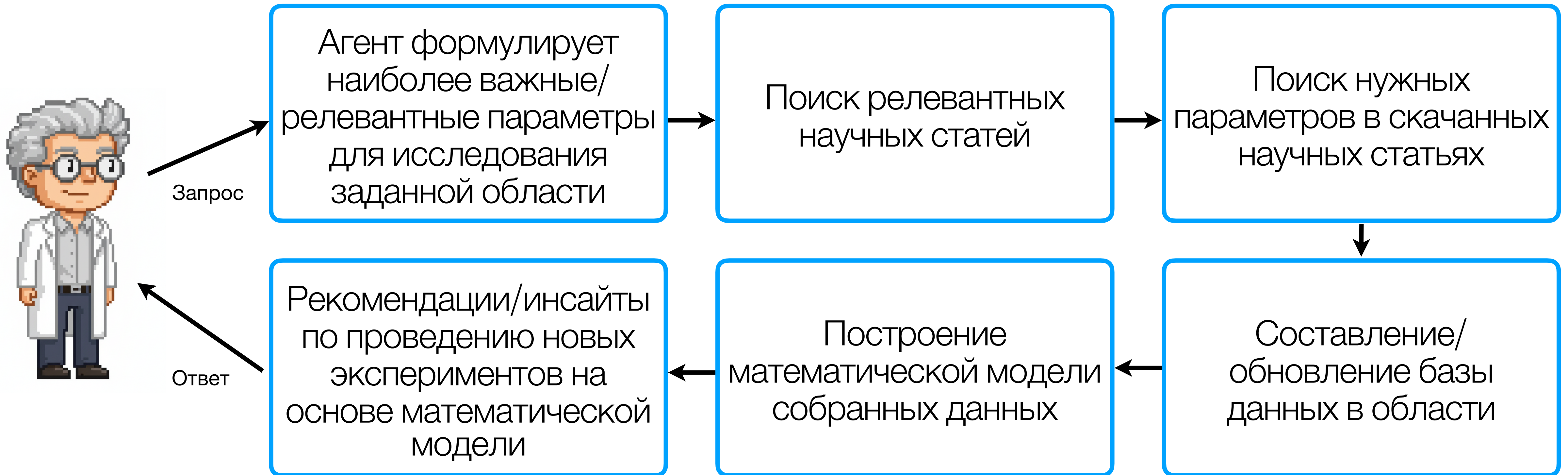
Поиск нужных параметров в скачанных научных статьях

Рекомендации/инсайты по проведению новых экспериментов на основе математической модели

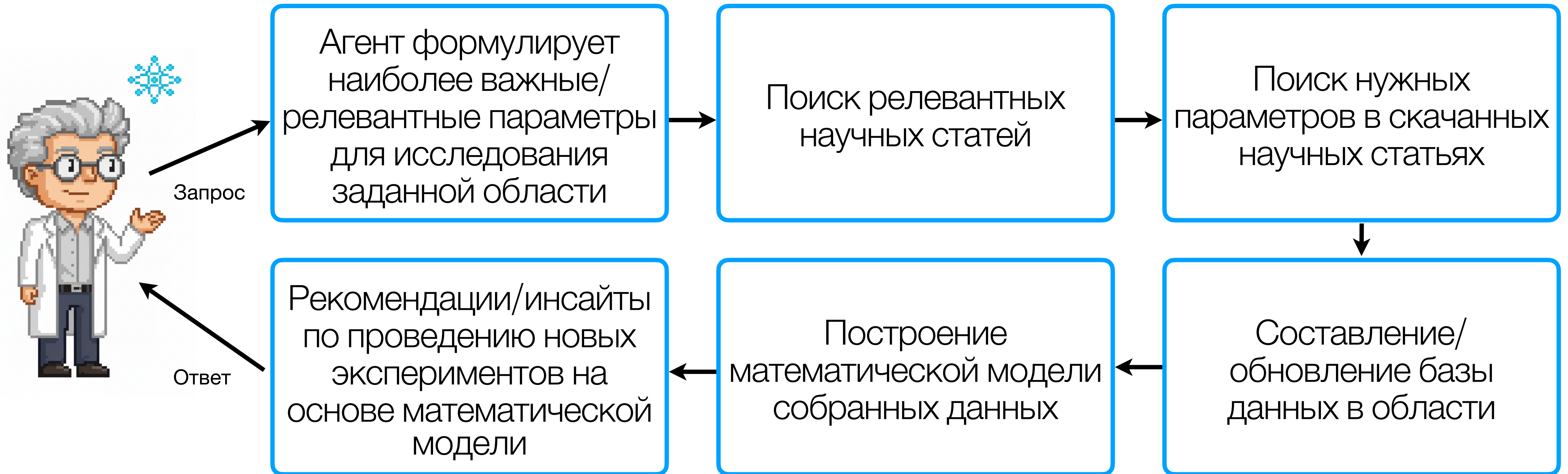
Построение математической модели собранных данных

Составление/ обновление базы данных в области

Data-driven science



Data-driven science



Агентная наука в разных доменных областях

