DLS

ASR for low-resource

Кузьменко Андрей

ML-engineer, SBER

2025

Почему low-resource ASR сложная задача?

Компьютерное зрение

- Простая адаптация, масштаб — всё человечество, простой краудсорс для разметки данных (инвариантность у людей / мета домен)

Обработка естественного языка

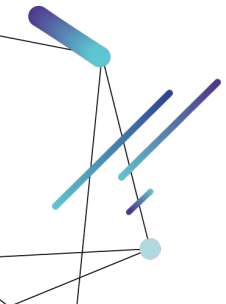
- Средняя адаптация, масштаб — все носители конкретного языка, простой краудсорс (инвариантность в рамках одного языка / мета домена)

Обработка речи

- Сложная адаптация, масштаб — носители языка и специализированные домены, дорогая разметка, сложный краудсорс

DLS

Модели и обучение



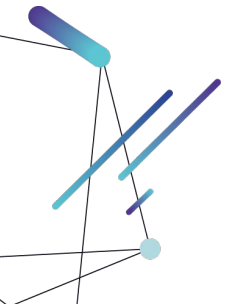
2025

Подходы обучения для low-resource

Transfer learning (TL) - Адаптация моделей, обученных на больших датасетах (supervised), для переноса знаний на другие домены (единицы и десятки часов) или родственные языки (десятки часов).

- Необходима модель родственная обученная модель (близкие языки/домены)

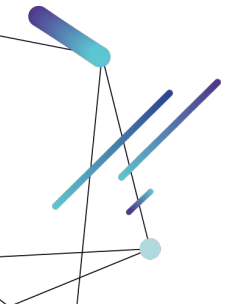
+ быстрые эксперименты



Подходы обучения для low-resource

Self-supervised Learning (SSL) Обучение на неразмеченных данных (десятки сотни тысяч часов),(от десятков до сотен часов).

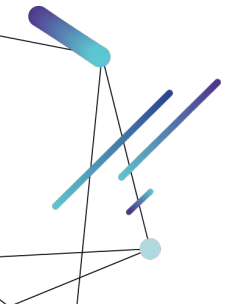
- большое количество неразмеченных данных для SSL
- большое количество вычислительных ресурсов
- длительные эксперименты
+ лучшее масштабирование от данных



Подходы обучения для low-resource

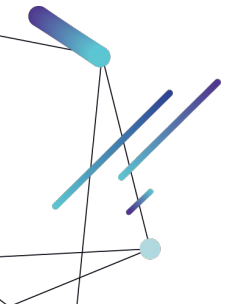
Комбинированный подход (SSL+TL) Предобучение кодировщика, обучение задачи распознавания речи, удаление избыточных доменов

- большое количество неразмеченных данных для SSL
- большое количество вычислительных ресурсов
- длительные эксперименты
+ потенциально более робастная модель для out of domain



DLS

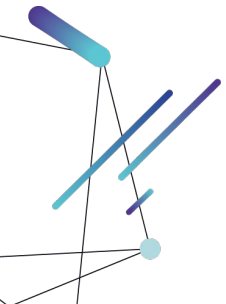
Transfer learning



2025

DLS

Attention-based Encoder Decoder (AED)



2025

Transfer learning (TL): AED

Multitask training data (680k hours)

English transcription

"Ask not what your country can do for ⋯"

Ask not what your country can do for ⋯

Any-to-English speech translation

"El rápido zorro marrón salta sobre ⋯"

The quick brown fox jumps over ⋯

Non-English transcription

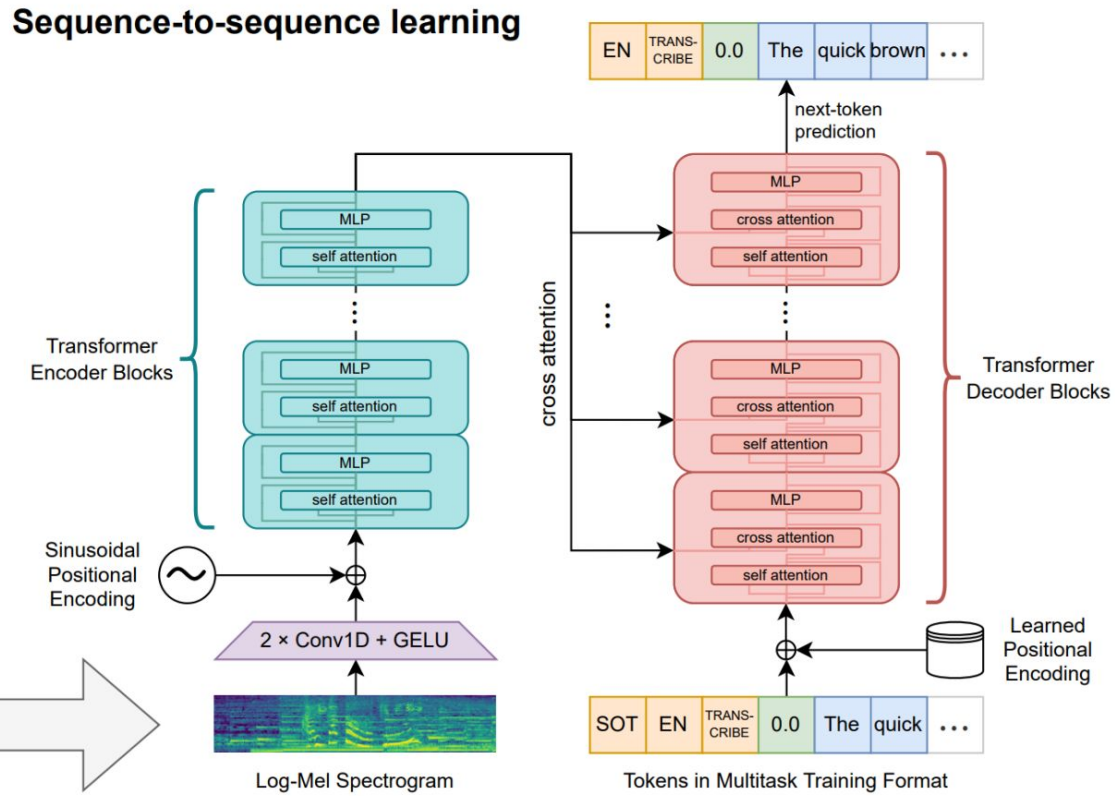
"언덕 위에 올라 내려다보면 너무나 넓고 넓은 ⋯"

언덕 위에 올라 내려다보면 너무나 넓고 넓은 ⋯

No speech

(background music playing)

∅



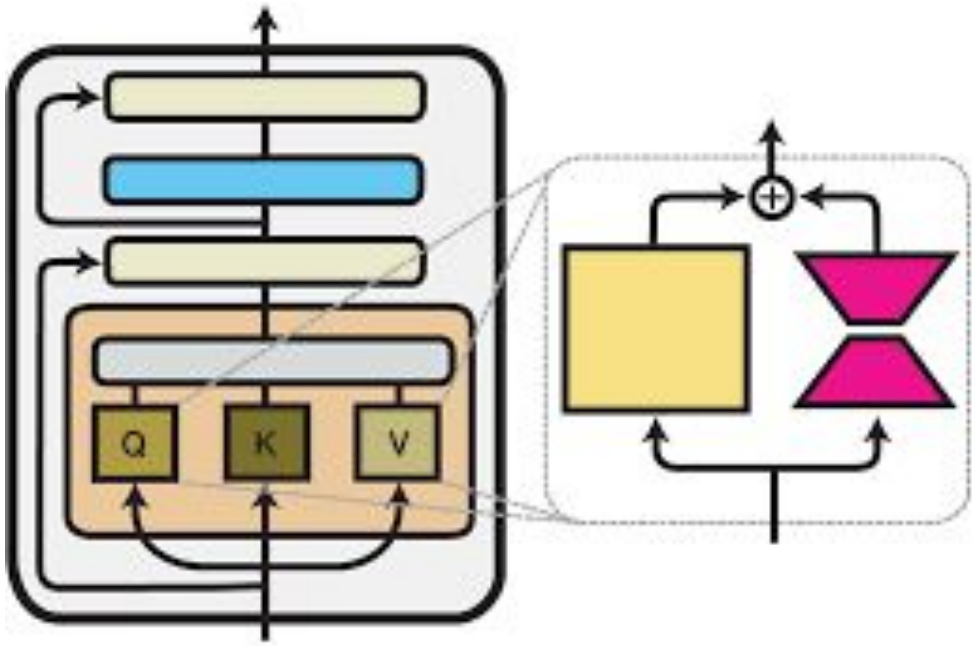
2025

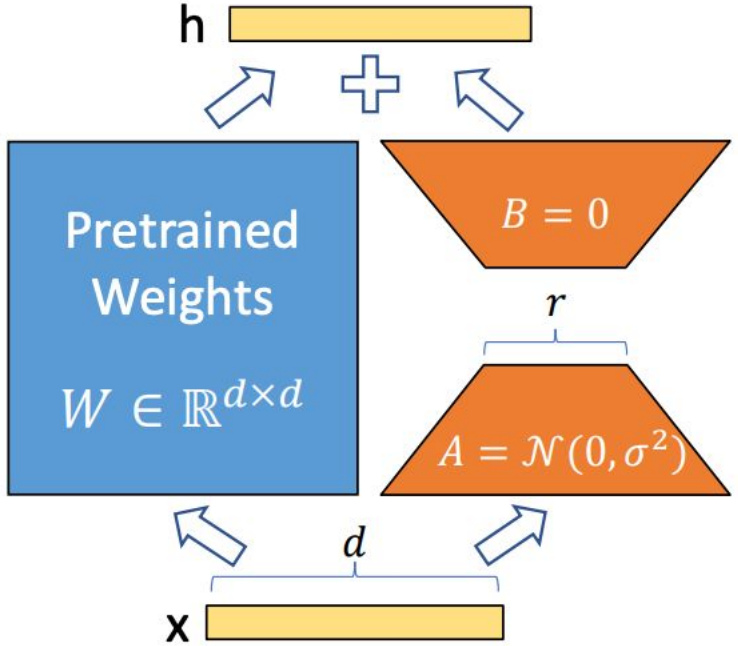
AED: finetune, преимущества и недостатки

+ Максимальная производительность
+ Простота реализации
+ Устоявшаяся практика

- Катастрофическая забывчивость
- Вычислительная дороговизна по памяти (при обучении)
- Быстрое переобучение / большое время обучения под целевой домен (при использовании родственных данных)

AED: PEFT


Q
K
V

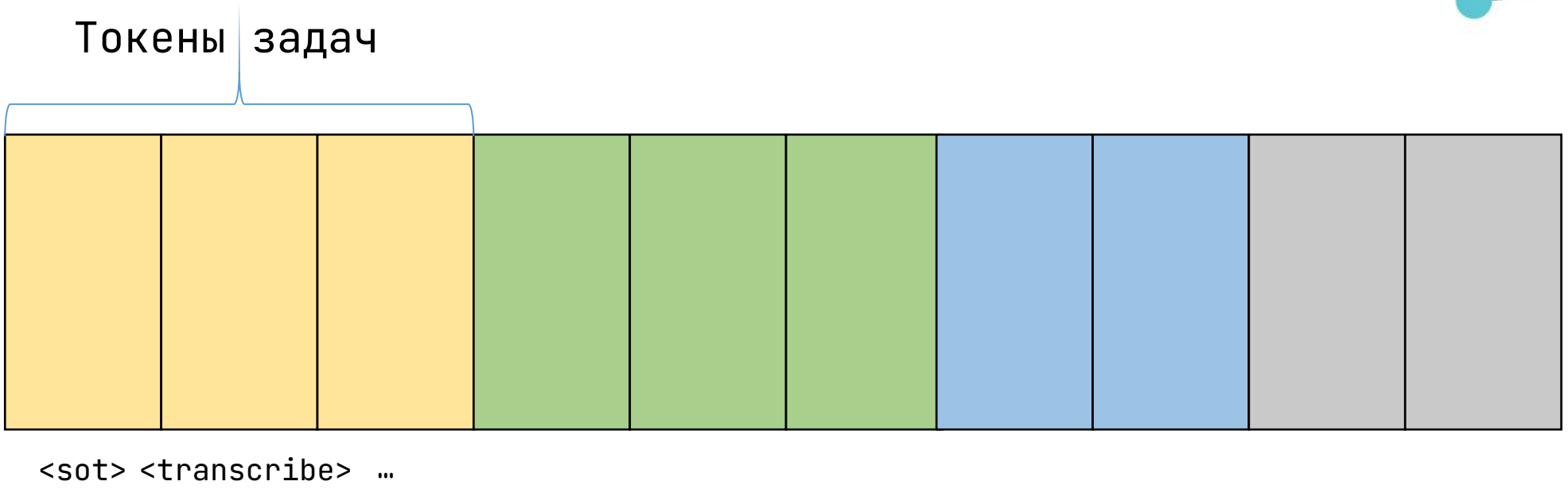

h
Pretrained Weights
$W \in \mathbb{R}^{d \times d}$
$B = 0$
r
$A = \mathcal{N}(0, \sigma^2)$
d
x

AED: PEFT

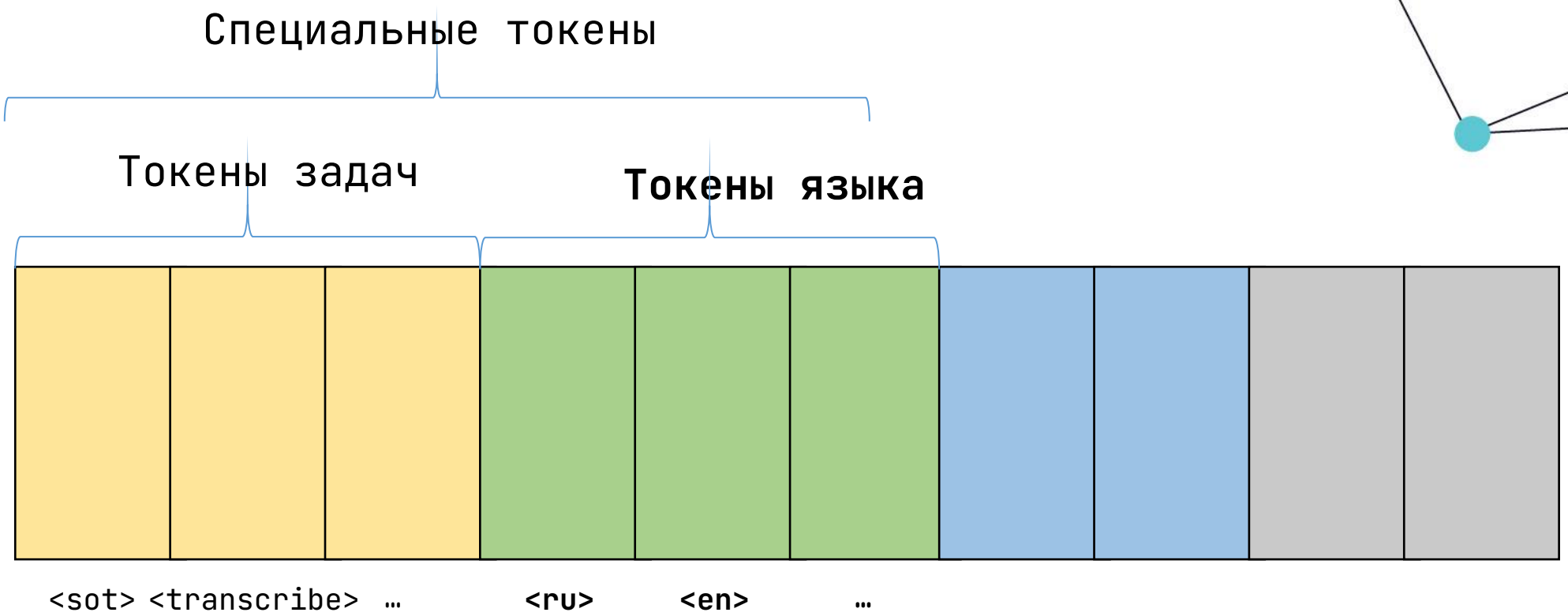
+ Эффективность по памяти (во время обучения)
+ Модульность и портативность (можно легко заменять адаптеры для базовой модели под различные домены)
+ Снижение риска переобучения
+ Сохранение общих знаний
+ Быстрое обучение под целевой домен (можем ставить большой lr)

- Потенциально менее низкое качество чем у finetune
- Сложность имплементации (хотя популярные фреймворков поддерживают из коробки)
- Дополнительные затраты на инференс

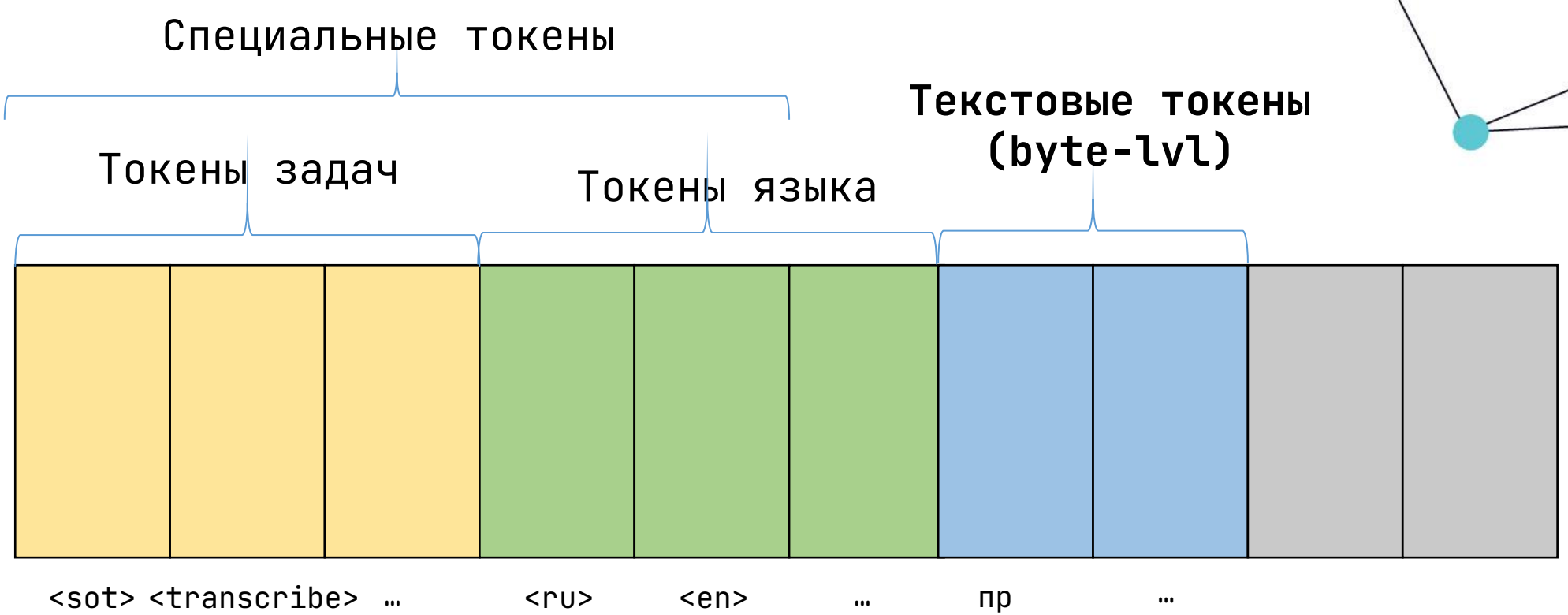
AED (whisper): Словарь



AED (whisper): Словарь

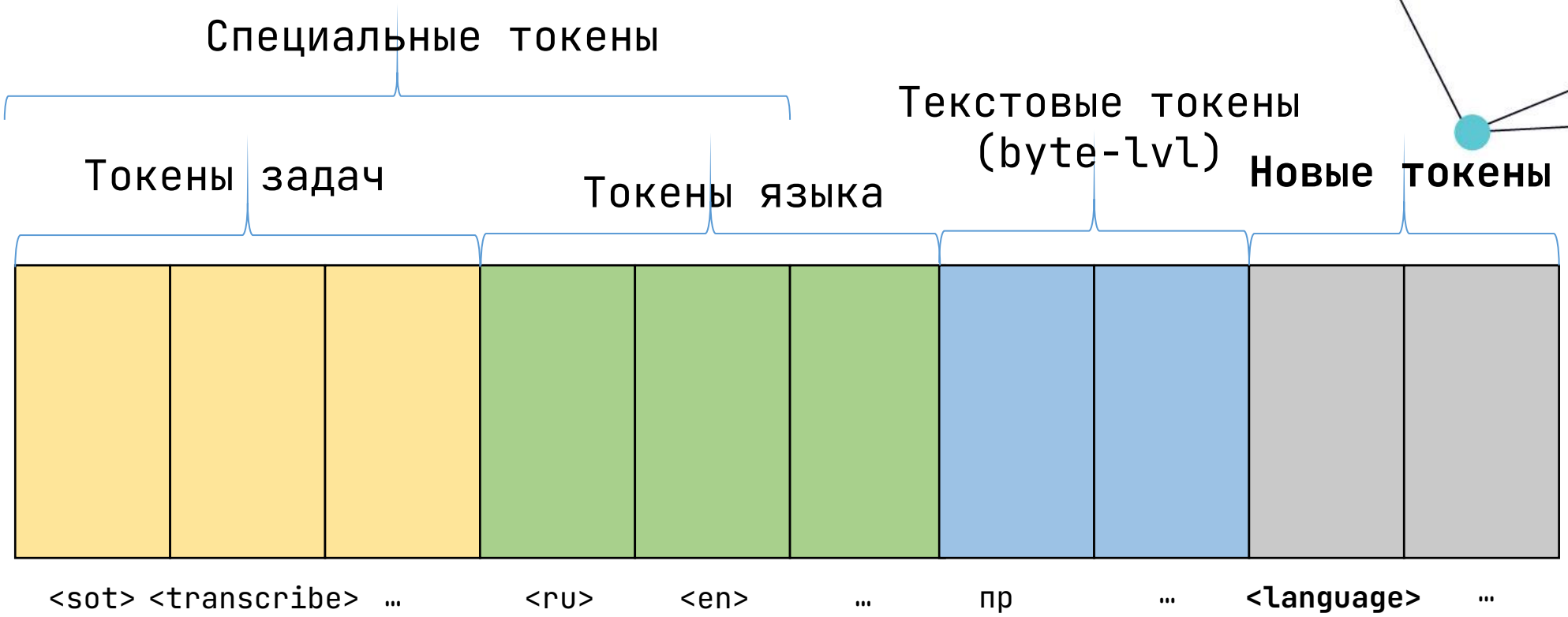


AED (whisper): Словарь



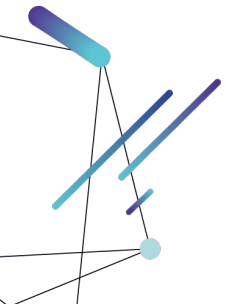
2025

AED (whisper): Словарь



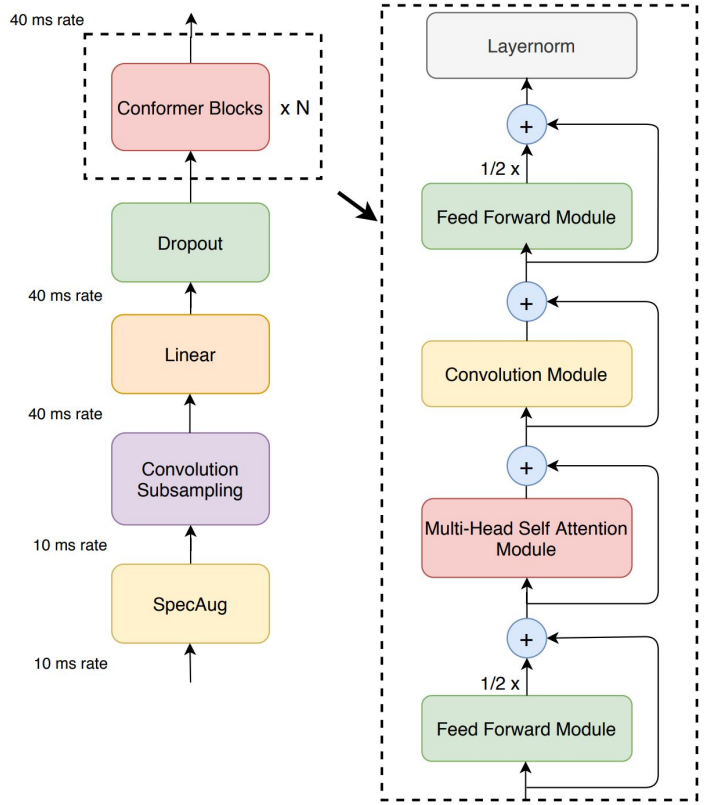
DLS

Connectionist Temporal Classification (CTC)

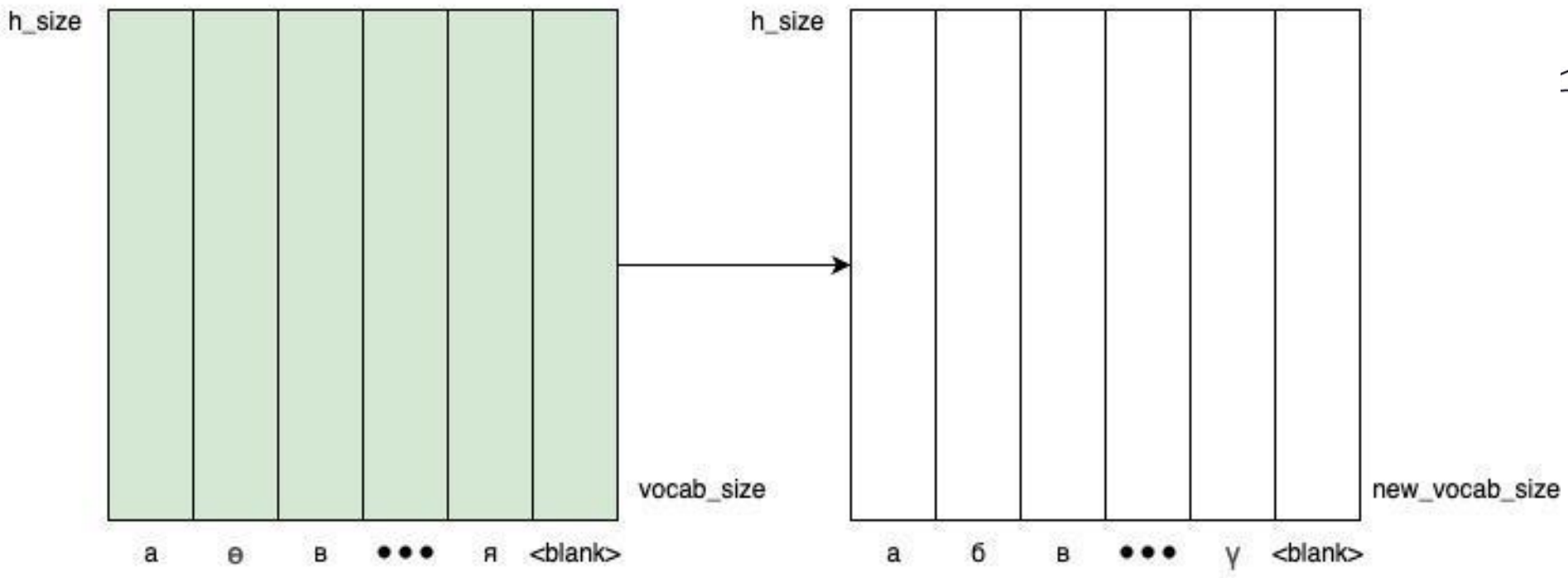


2025

CTC Conformer: Архитектура

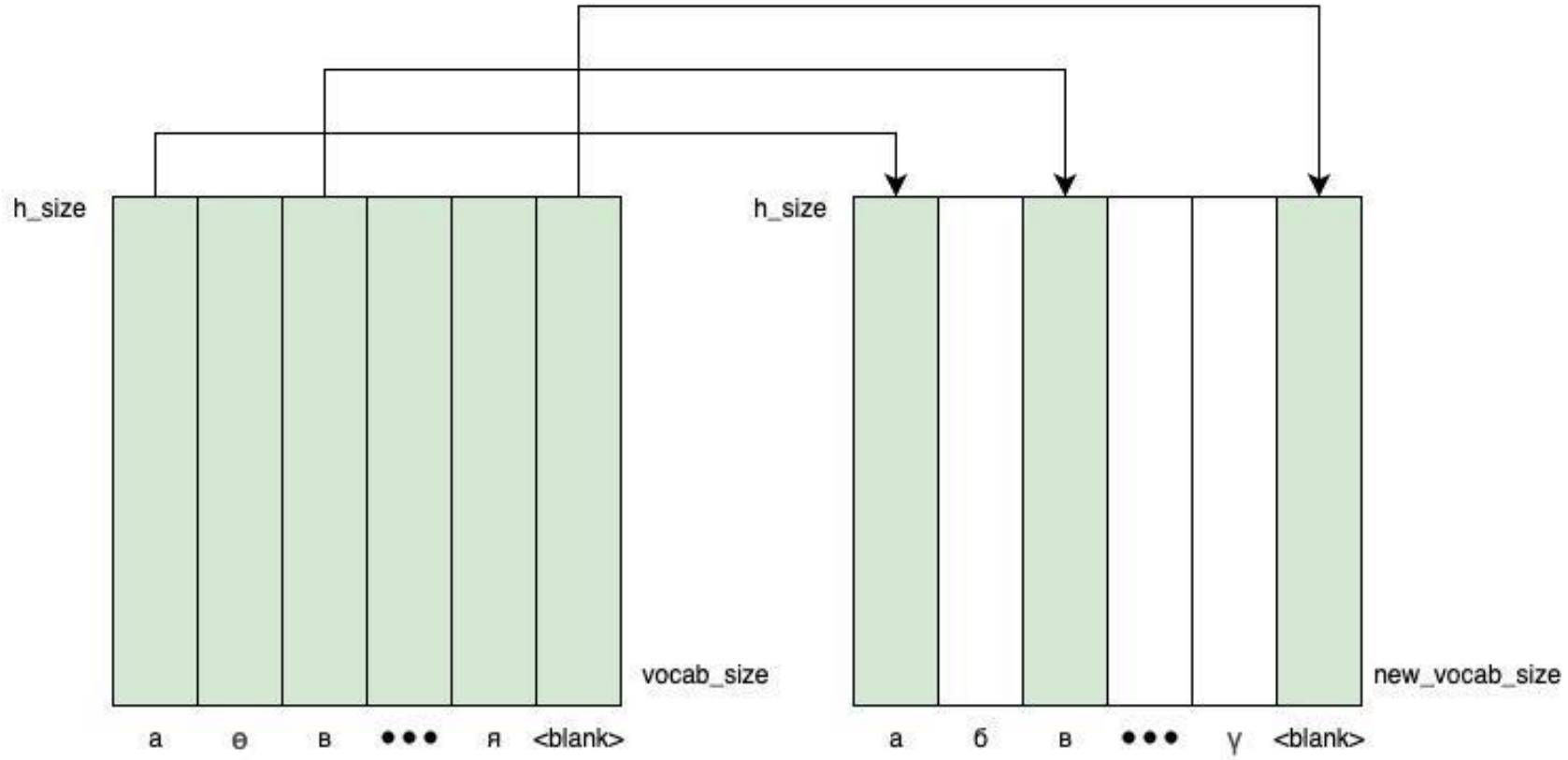


CTC conformer: перенос знаний на новый язык (кириллица->латиница)



2025

CTC conformer : перенос знаний на родственный язык

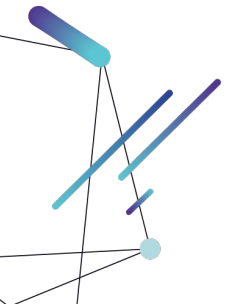


High resource decoder Low resource decoder

2025

DLS

Self-Supervised Learning



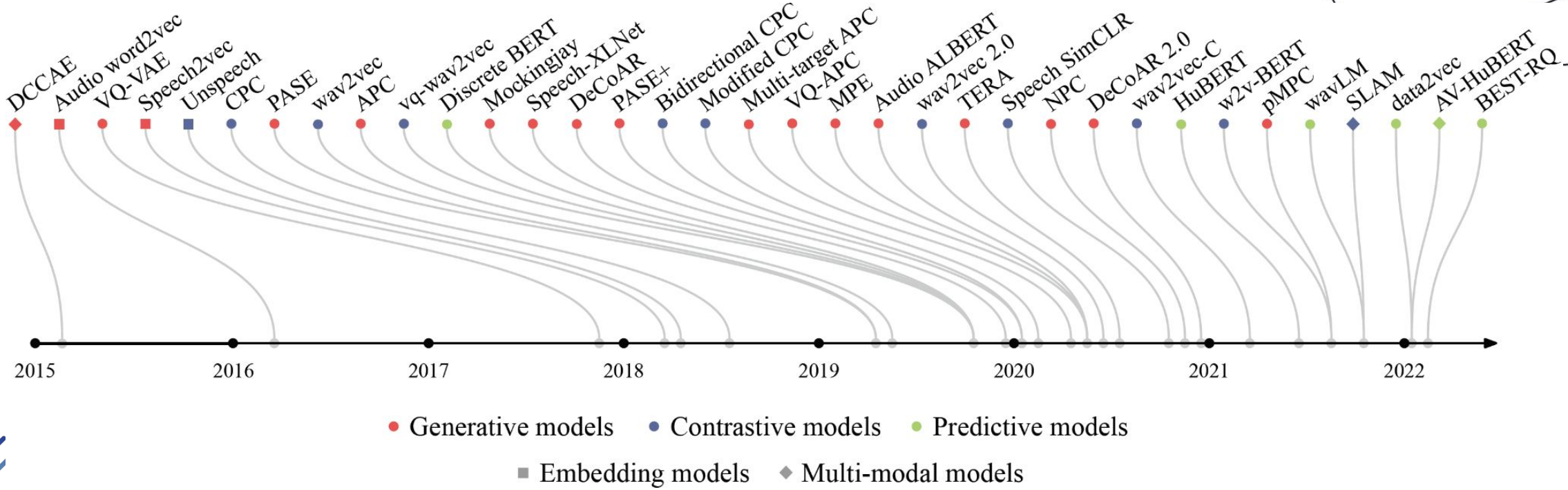
2025

SSL vs TL

+ Потенциально более высокое качество
+ Чувствительные данные

- Необходимо обрабатывать огромные объемы неразмеченных данных (высокие затраты на хранение, загрузку и тд)
- Наличие большого доменного корпуса неразмеченной речи
- Сложность имплементации:
  - Необходима эффективная работа загрузчиков (dataloader)
  - Отсутствие прямого критерия качества
  - Тяжело анализировать результат

Self-supervised learning for Speech



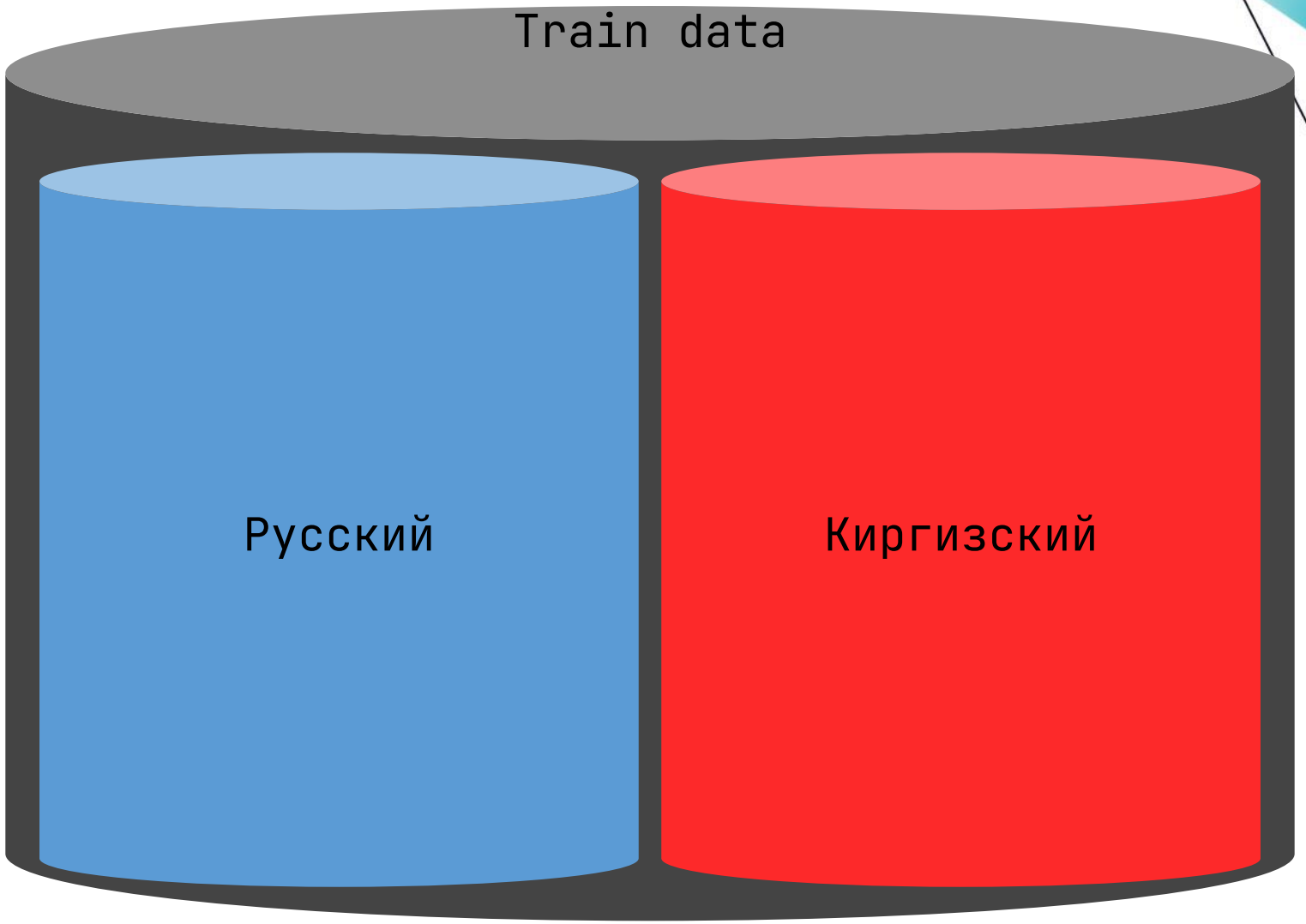
2025

DLS

Данные

2025

Базовый принцип

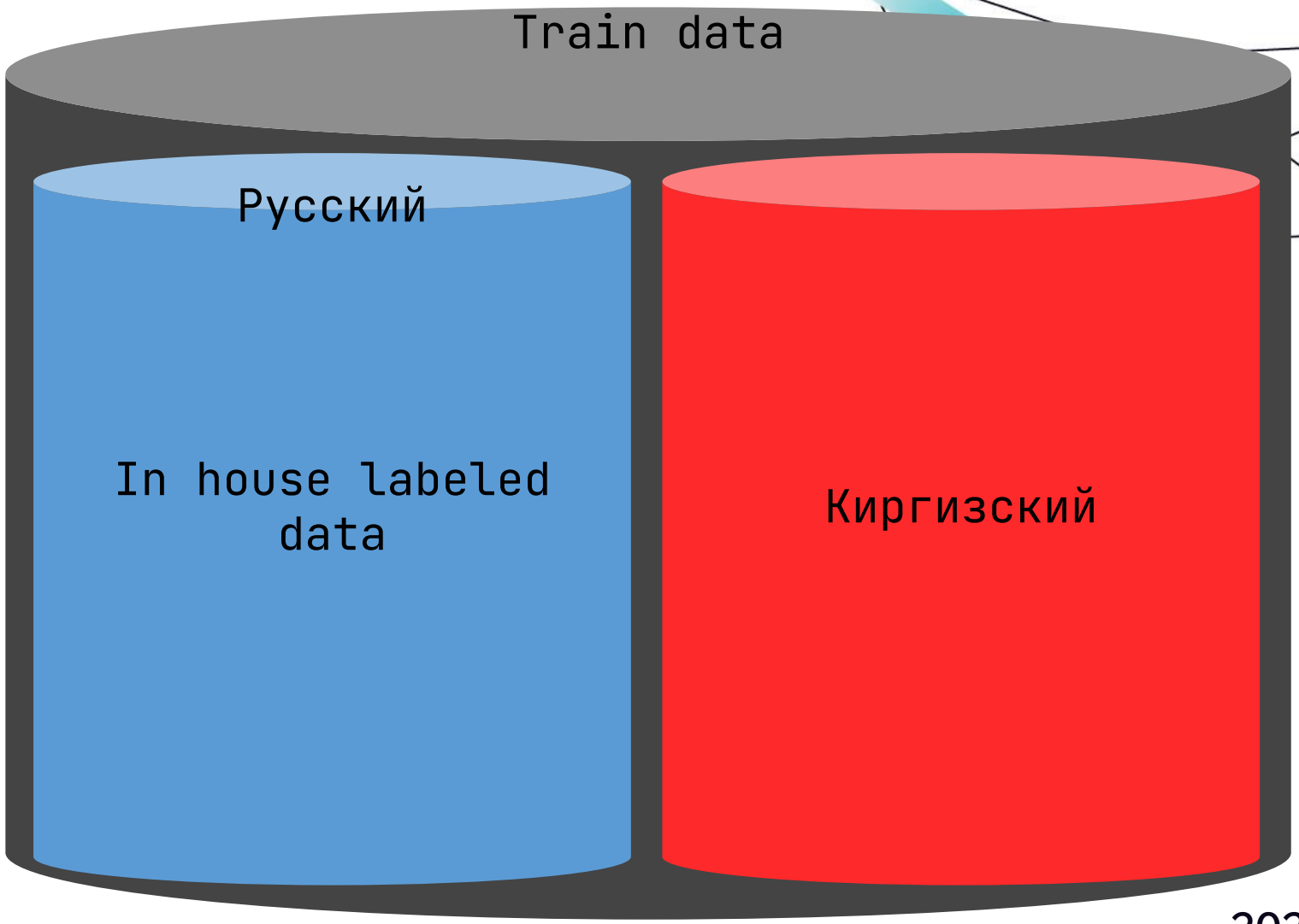


Данные

Русский (high resource)

- In house labeled data

Киргизский (low resource)



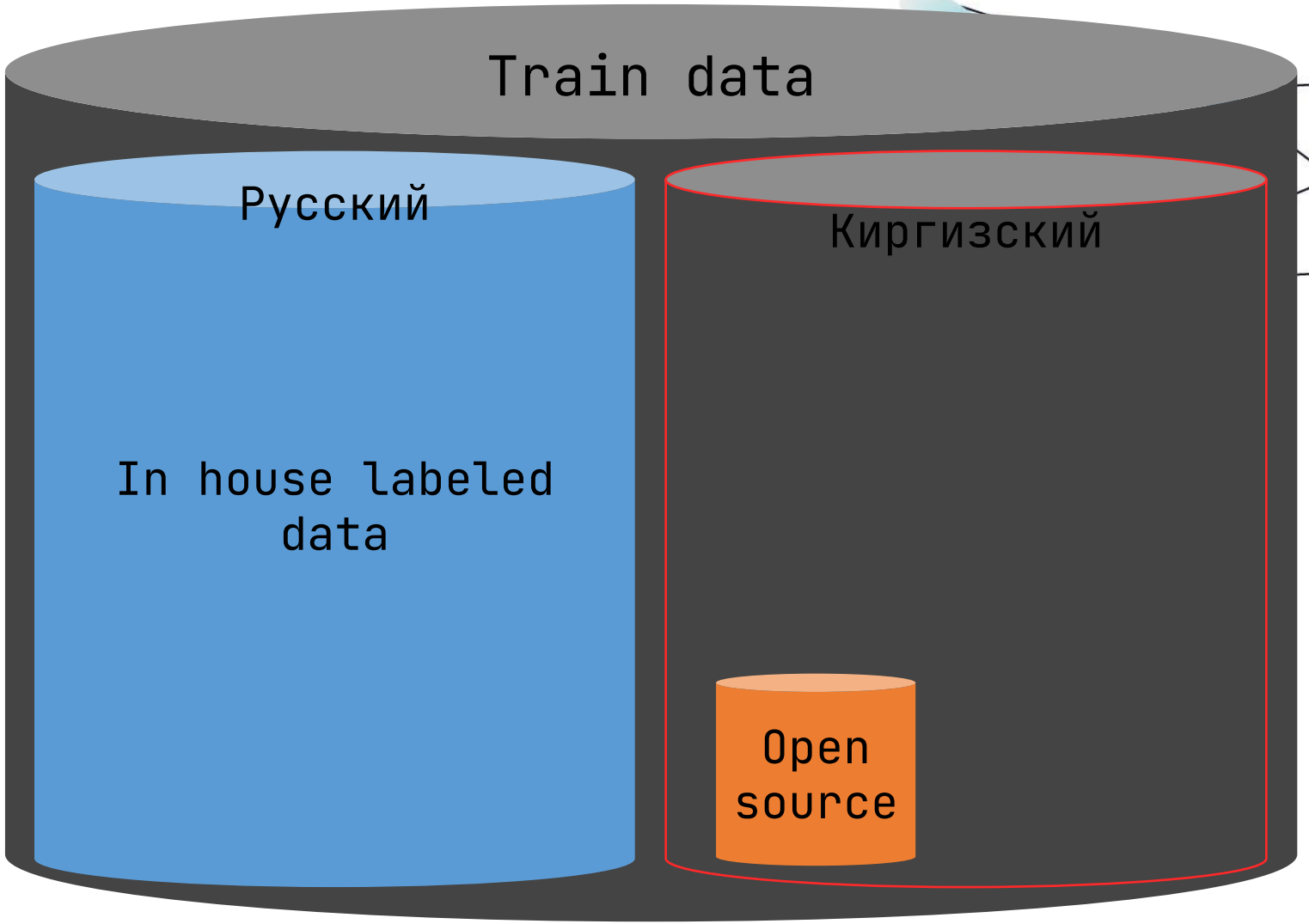
Open source

Русский (high resource)

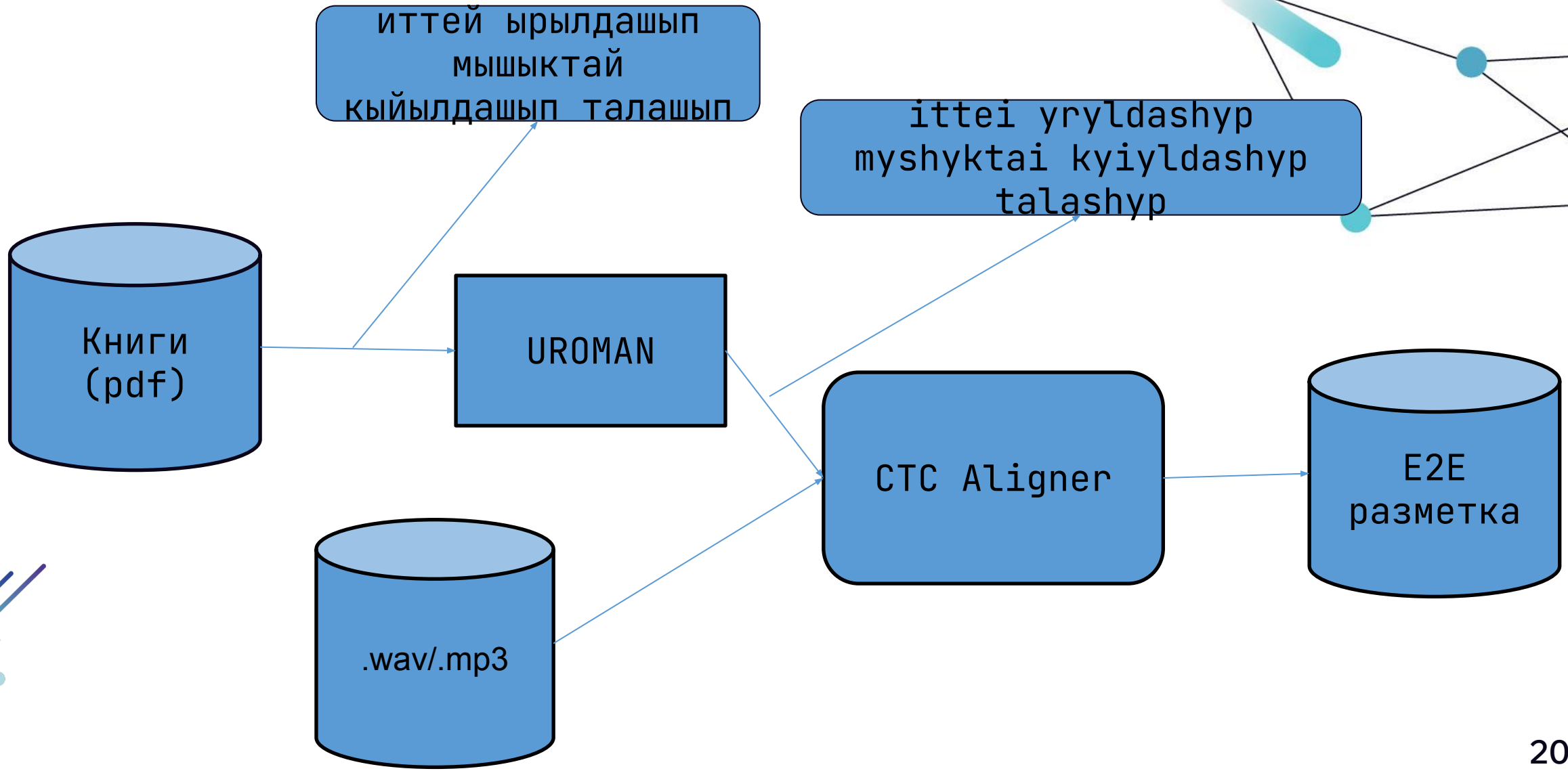
- In house labeled data

Киргизский (low resource)

- Open source



CTC-Forced alignment



2025

CTC-Forced alignment: алгоритм

- **Входные данные:**
  - Аудиозапись, преобразованная в последовательность акустических признаков (фреймов) X = [x1, x2, ..., xT].
  - Транскрипция — последовательность слов или символов Y = [y1, y2, ..., yU].
- **Проход модели:** Акустические признаки X пропускаются через предварительно обученную CTC-модель. На выходе получается матрица вероятностей P размером T x N, где N — размер алфавита (включая специальный токен blank). Каждый элемент P(t, k) — это вероятность того, что на фрейме t модель предсказывает символ k
- **Сопоставление транскрипции:** Текстовая транскрипция Y преобразуется в последовательность меток L, совместимую с алфавитом модели **Выравнивание:** Алгоритм ищет **наиболее вероятное выравнивание** между последовательностью фреймов X и метками L.
- **Выходные данные:** Для каждого фрейма аудио присваивается метка символа (или blank), указывающая, какой сегмент транскрипции произносится в этот момент времени.

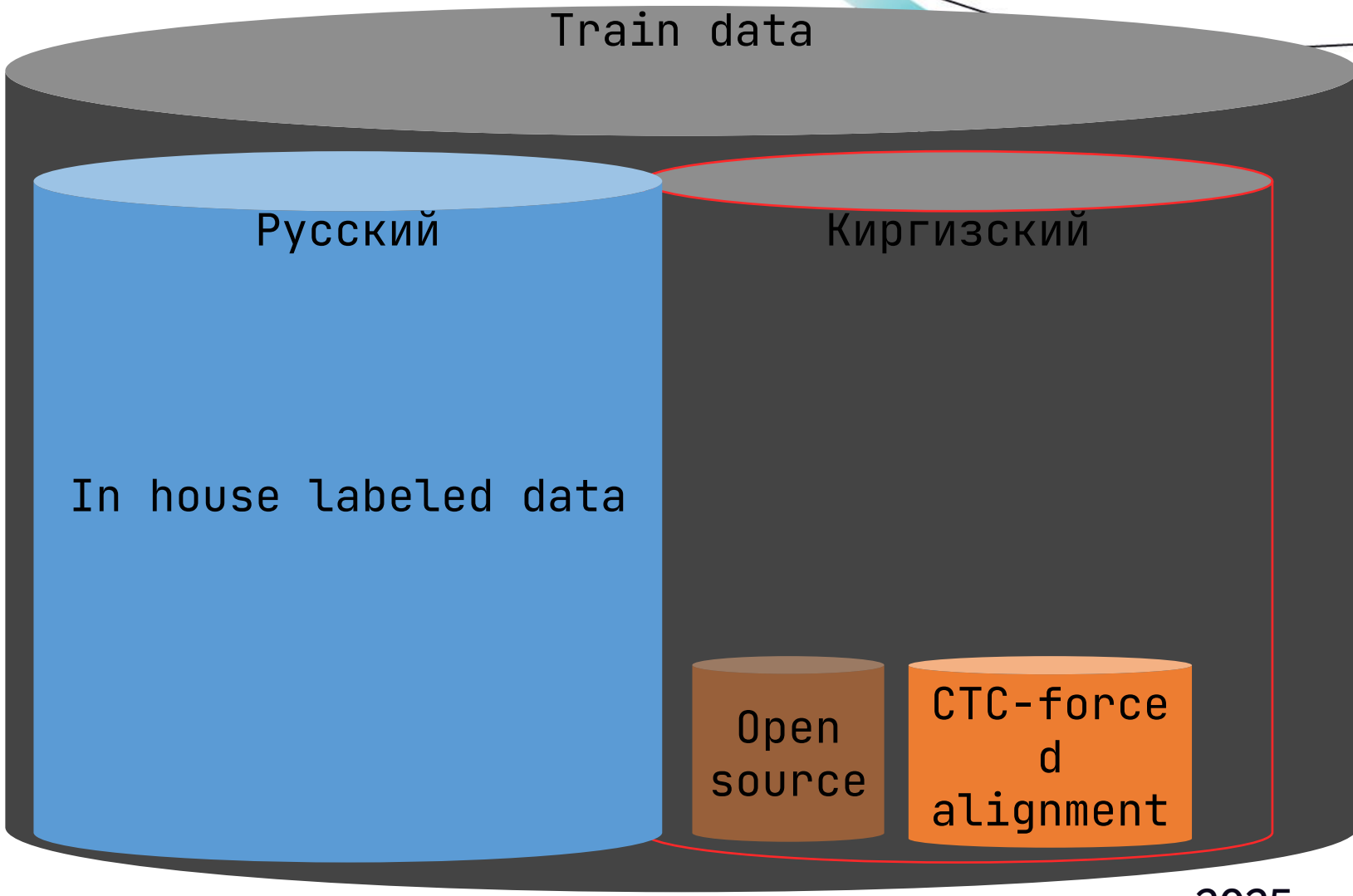
CTC-Forced alignment

Русский (high resource)

- In house labeled data

Киргизский (low resource)

- Open source
- CTC-forced alignment



Выводы, предположения

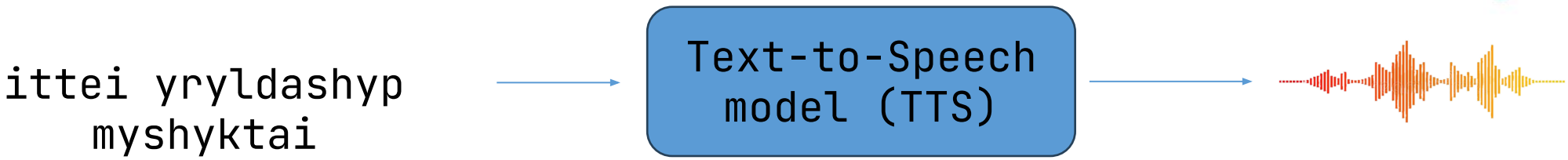
Звук

- Мало данных
- Out of domain речь
- «Качественная речь»: single speaker, нет прерываний, дефектов

Тексты

- Малый словарный запас
- Out of domain текст
- Отсутствие code switching (только киргизский)
- Отличные правила разметки (обработка слов

Синтетика



Псевдодоменная синтетика

какая сегодня погода

Machine Translation model (ru→ky)

бүгүн аба ырайы кандай

Text-to-Speech model (TTS)

2025

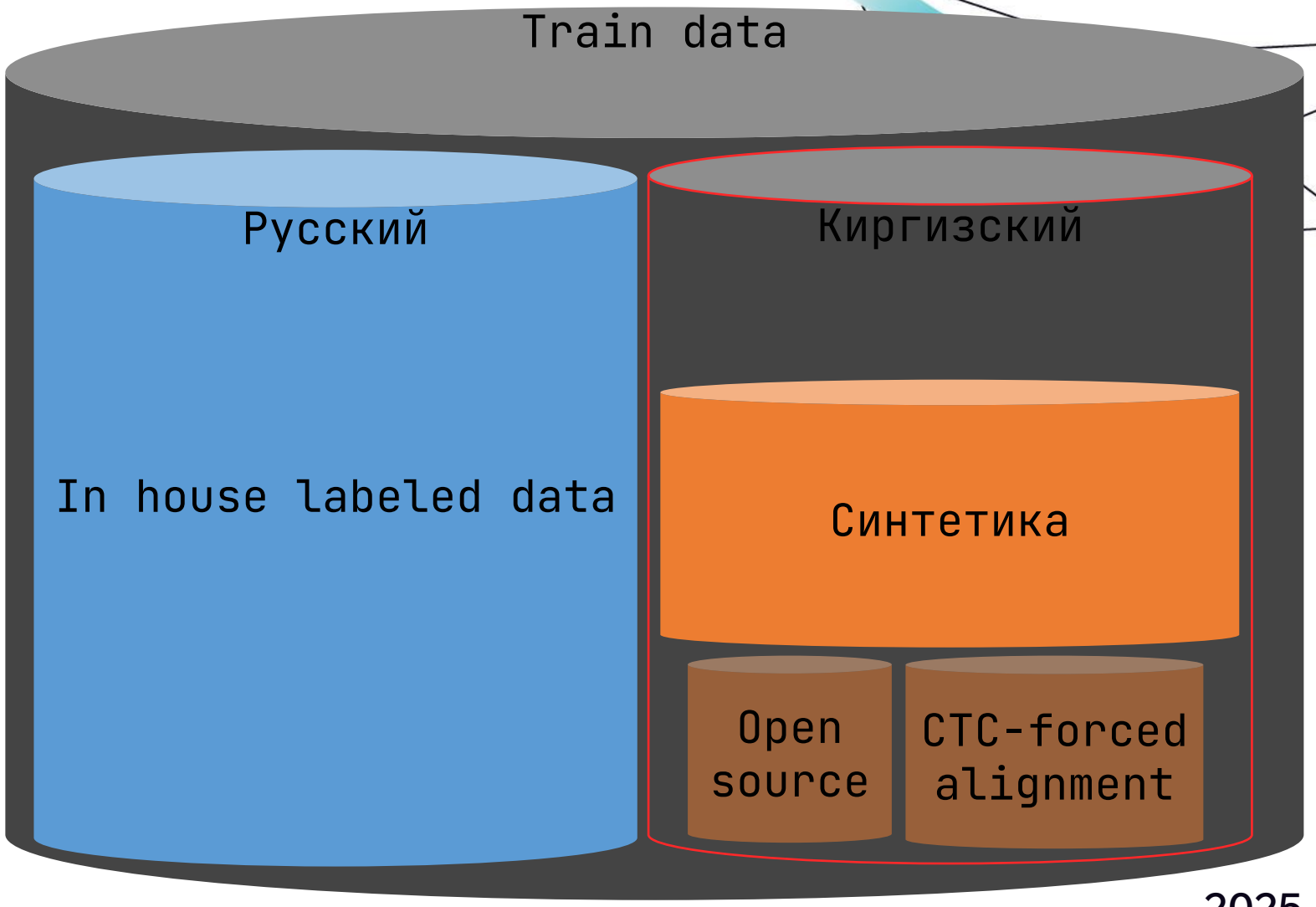
Данные: синтетика

Русский (high resource)

- In house labeled data

Киргизский (low resource)

- Open source
- CTC-forced alignment
- Синтетика



Выводы, предположения 2

Звук

- Мало данных ✓
- Out of domain речь
- «Качественная речь»: single speaker, нет прерываний, дефектов
- Однообразная речь

Тексты

- Малый словарный запас ✓
- Out of domain текст
- Отсутствие code switching (только киргизский)
- Отличные правила разметки (обработка слов паразитов, обрываний слов и тд)

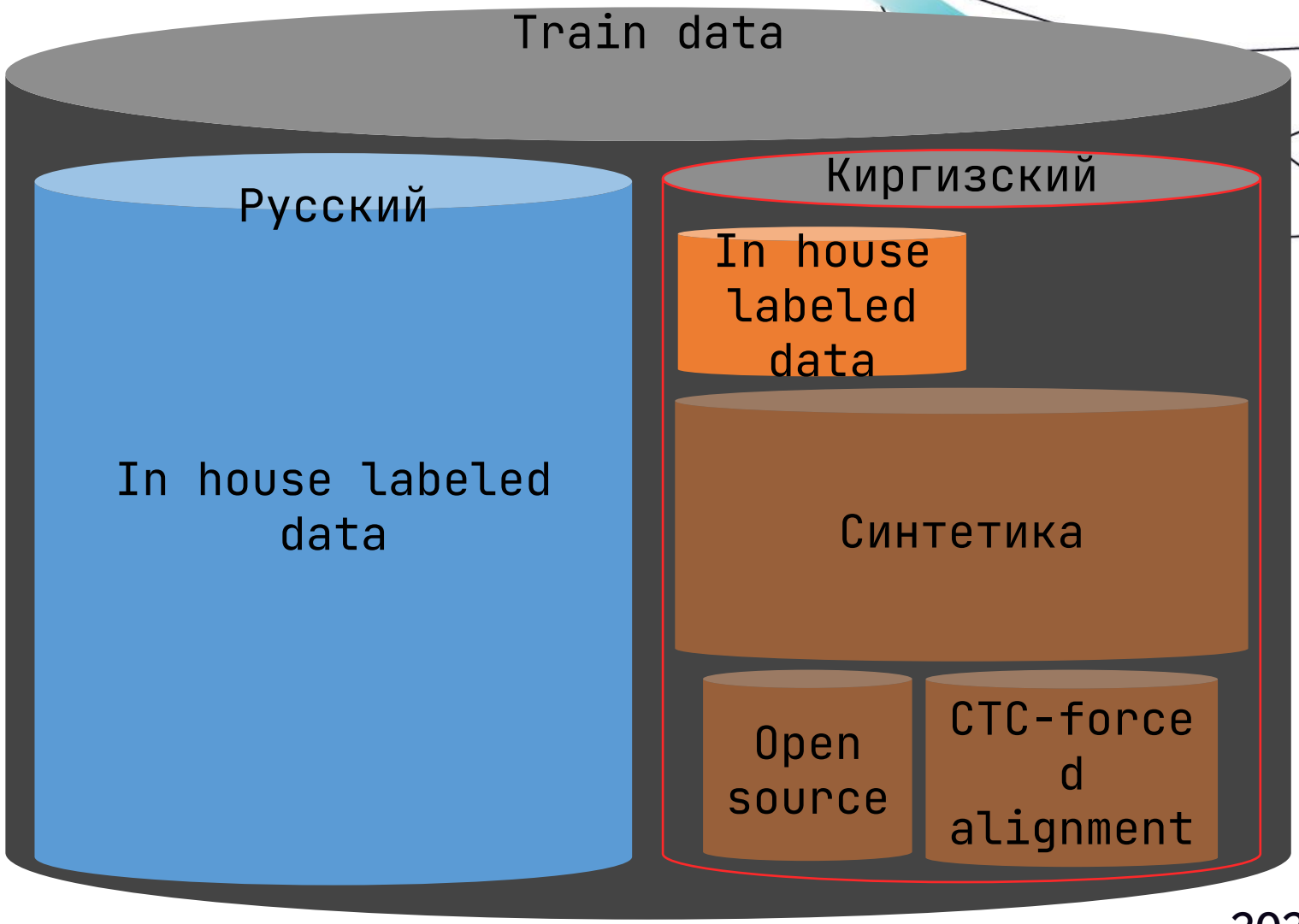
Данные

Русский (high resource)

- In house labeled data

Киргизский (low resource)

- Open source
- CTC-forced alignment
- Синтетика
- In house labeled data



Выводы, предположения 3

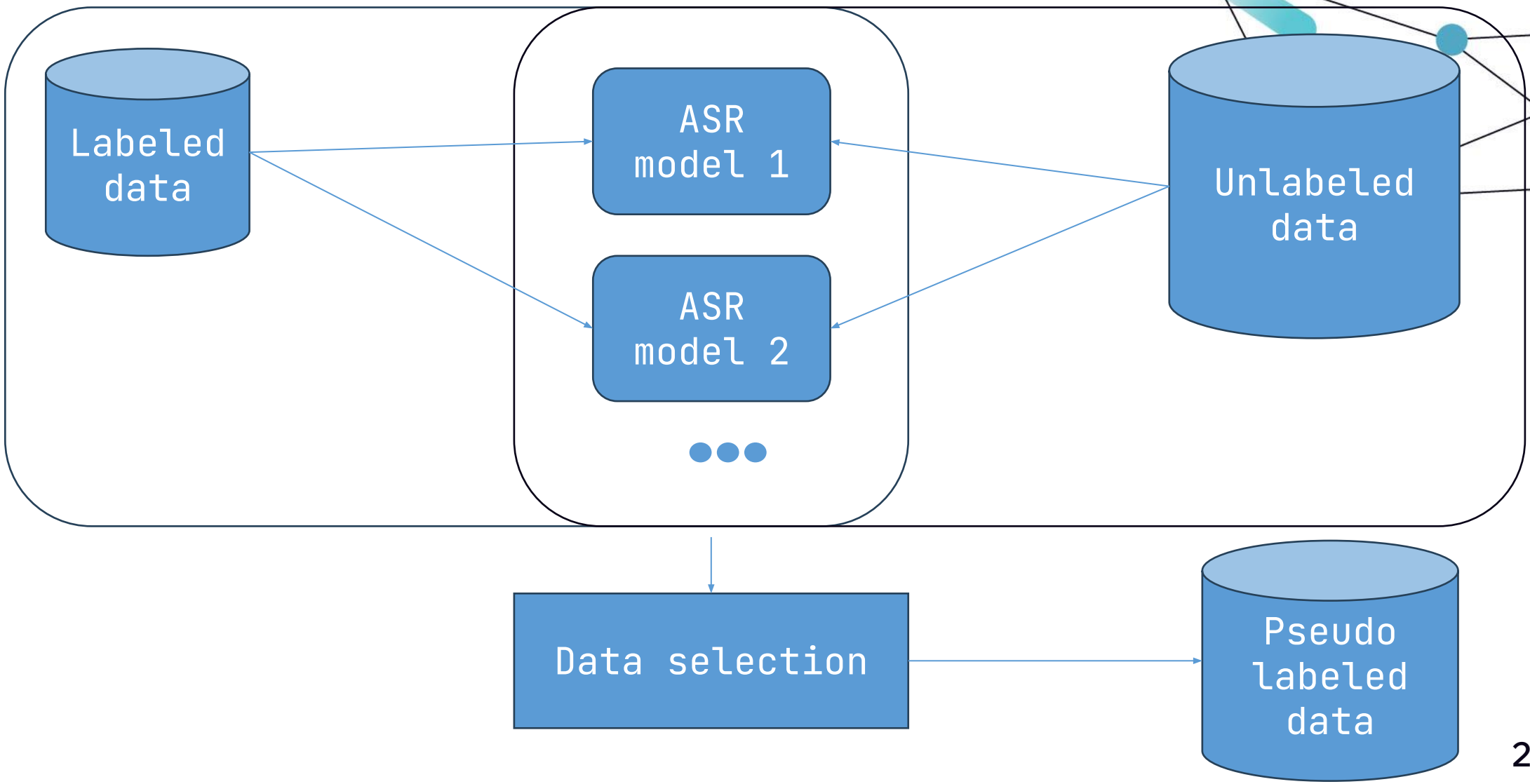
Звук

- Мало данных
- Out of domain речь
- «Качественная речь»: single speaker, нет прерываний, дефектов
- Однообразная речь
- Мало in domain речи

Тексты

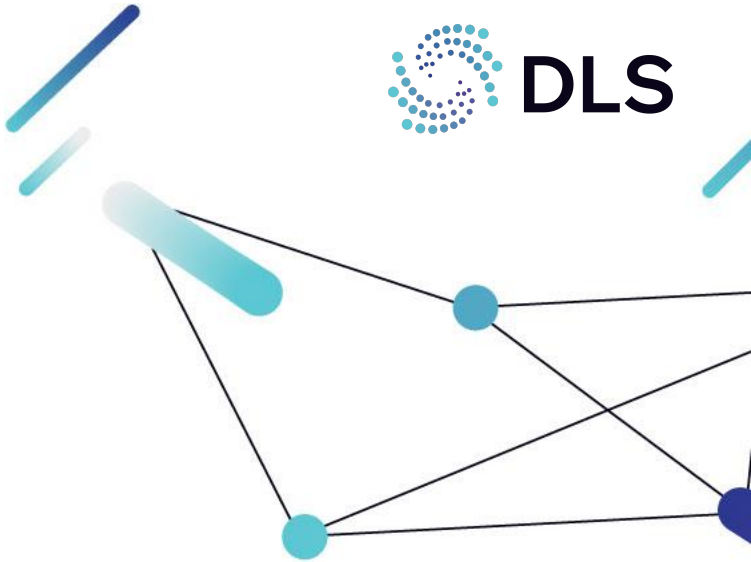
- Малый словарный запас
- Out of domain текст
- Отсутствие code switching (только киргизский)
- Отличные правила разметки (обработка слов паразитов, обрываний слов и тд)
- Мало in domain текстов

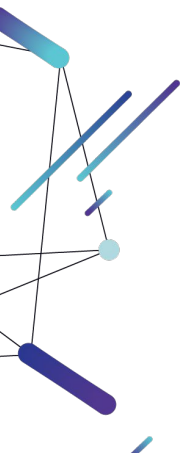
Pseudolabels: Pipeline



Pseudolabels

- Согласованность
- Rover
- Rescoring
- MWER Rescoring





DLS

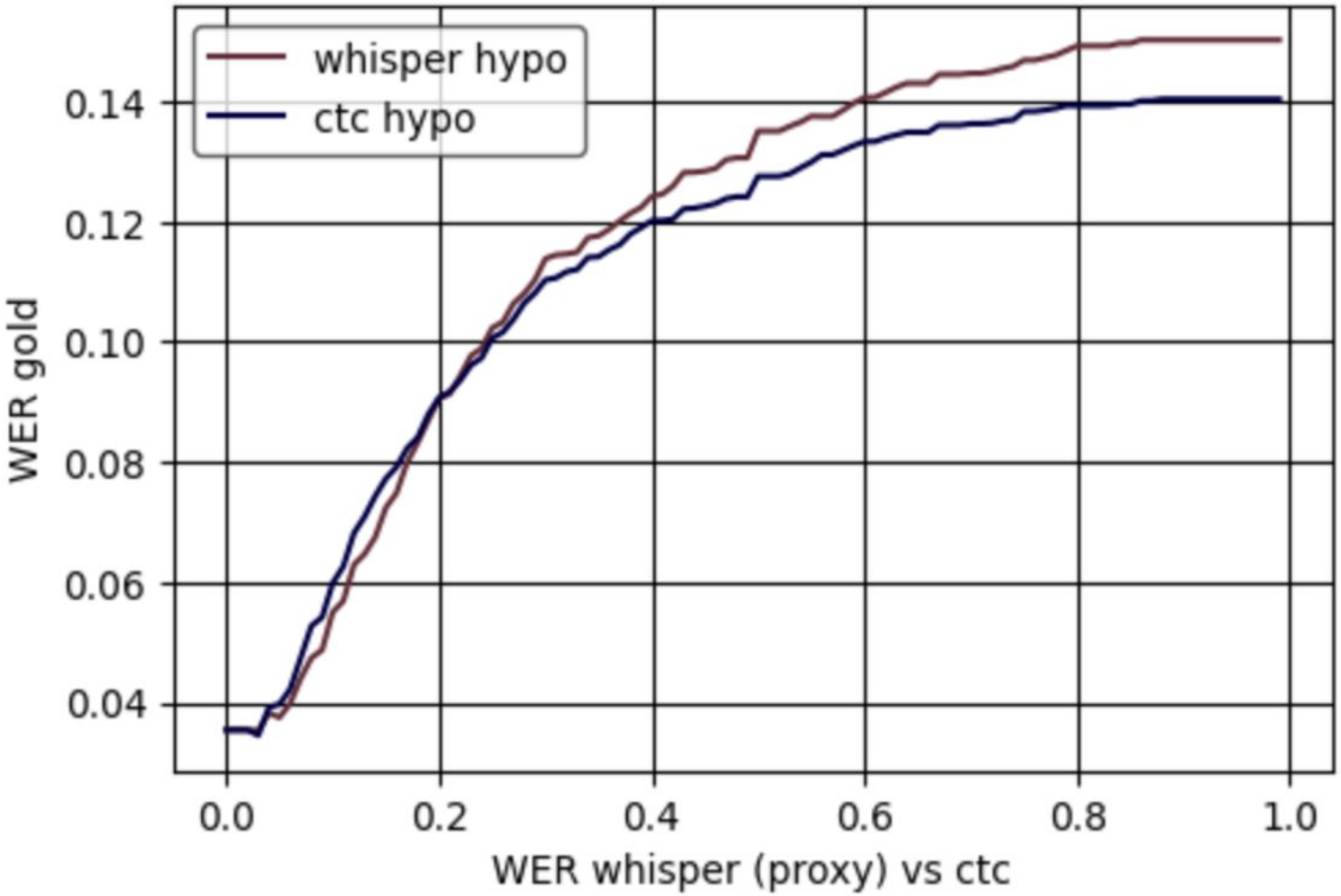
2025

DLS

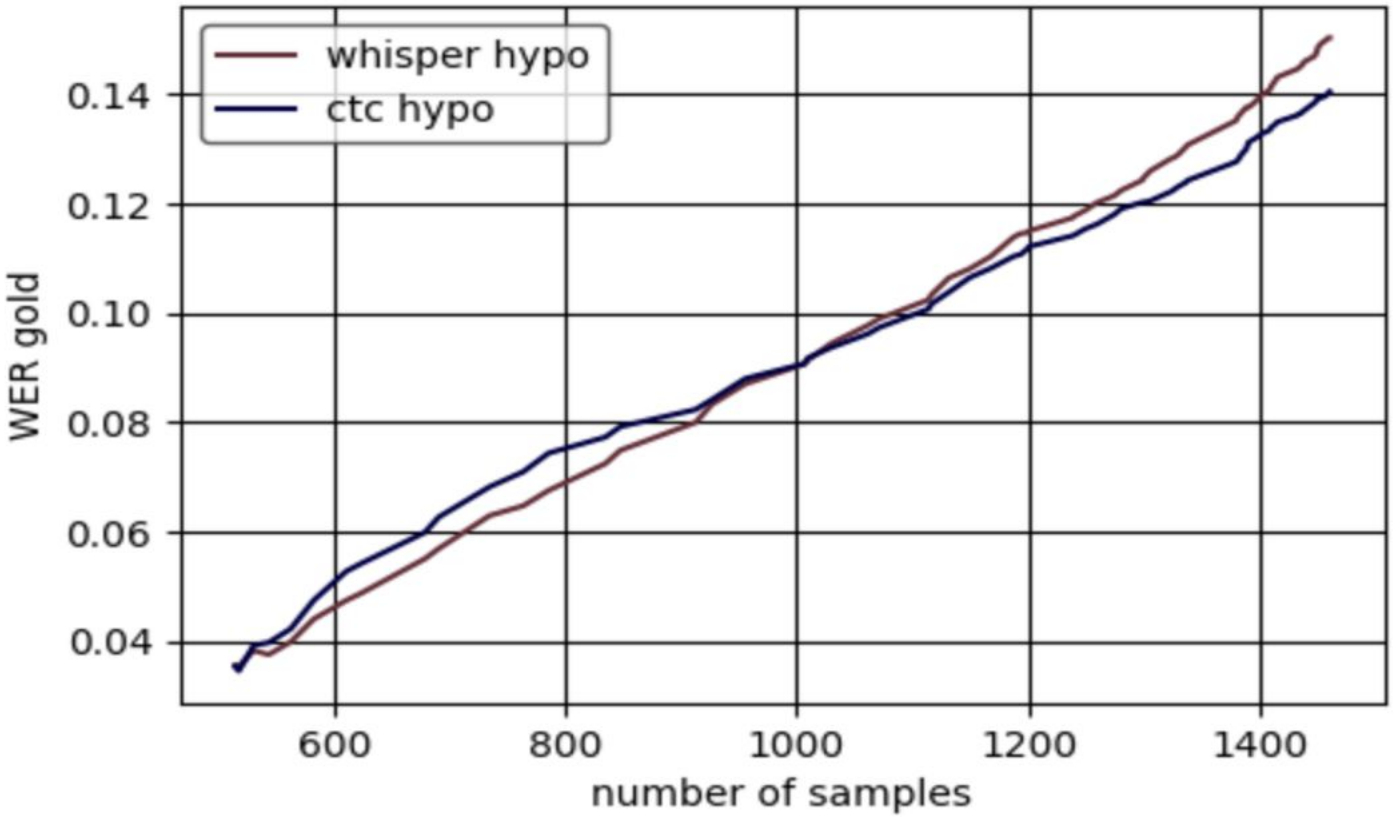
Согласованность

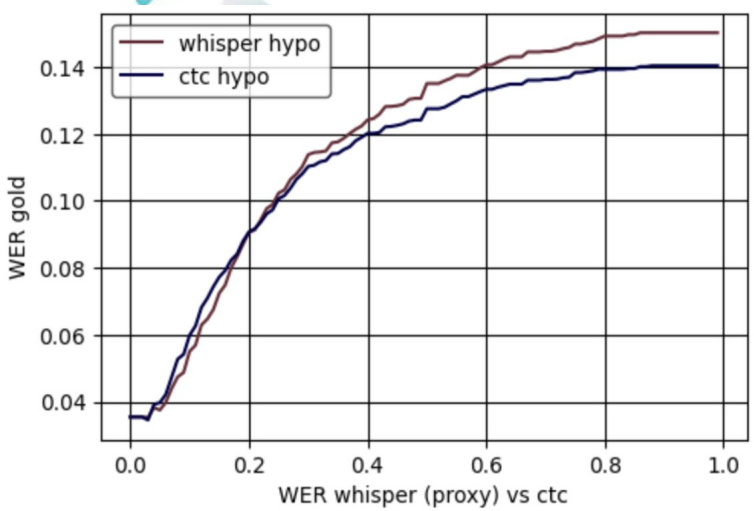
2025

Pseudolabels: Согласованность



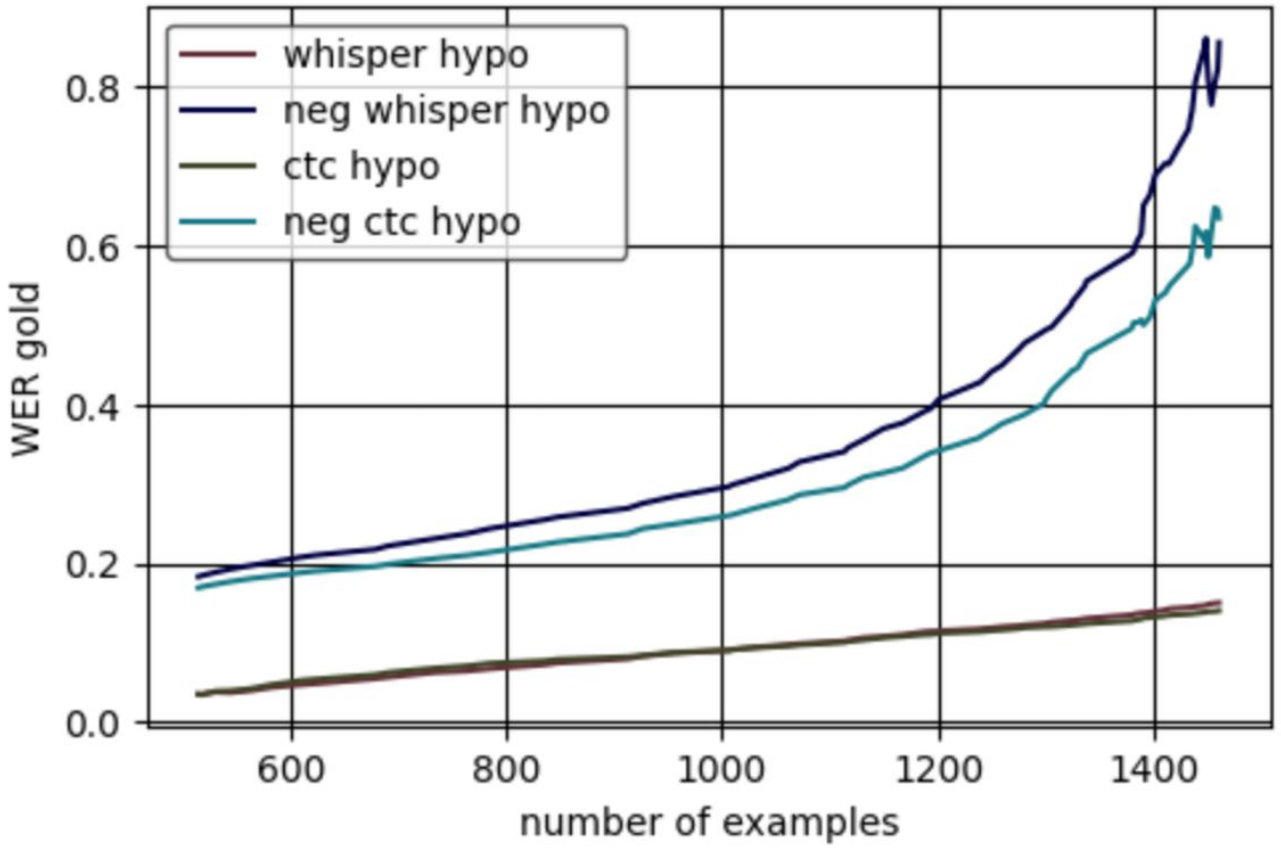
Pseudolabels: оценка объема подвыборки

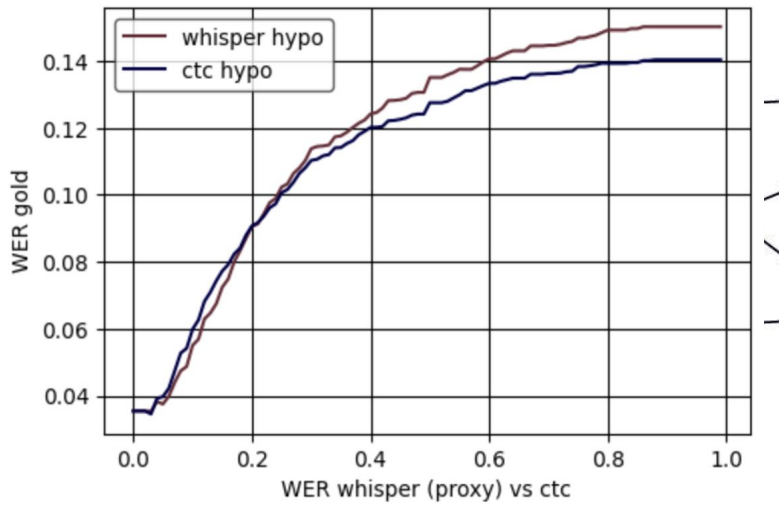


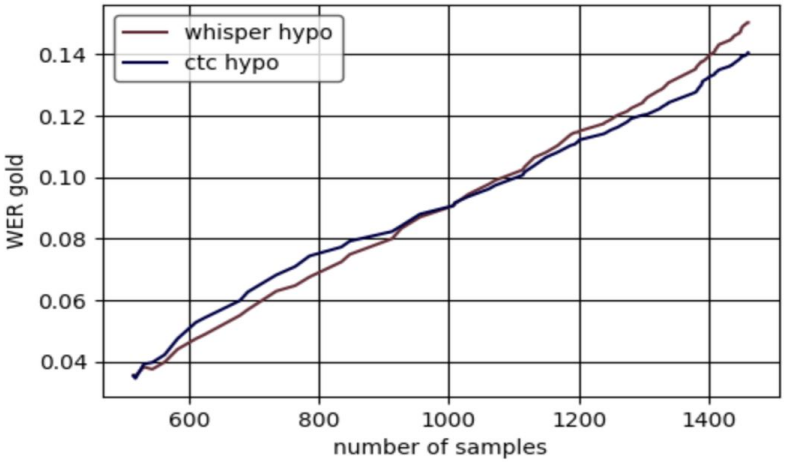


2025

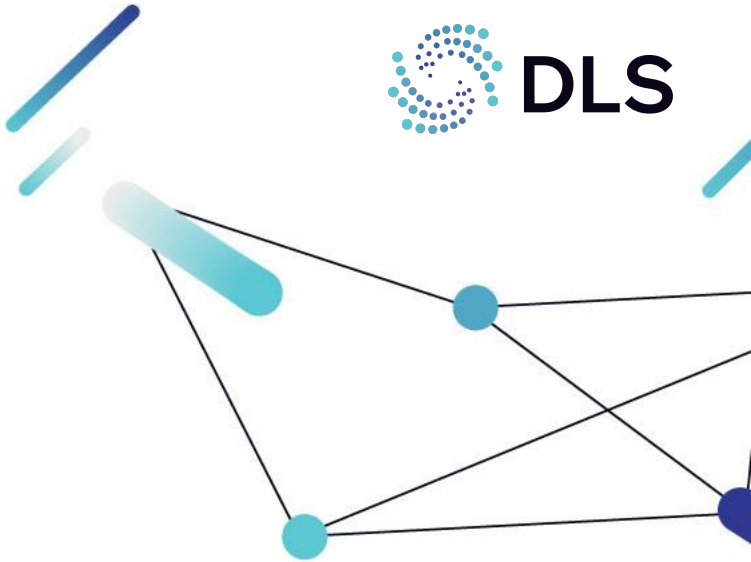
Pseudolabels: оценка объема подвыборки



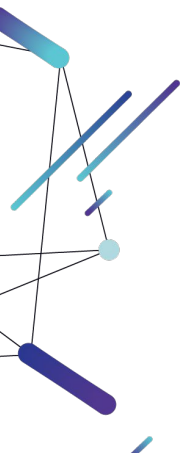




2025

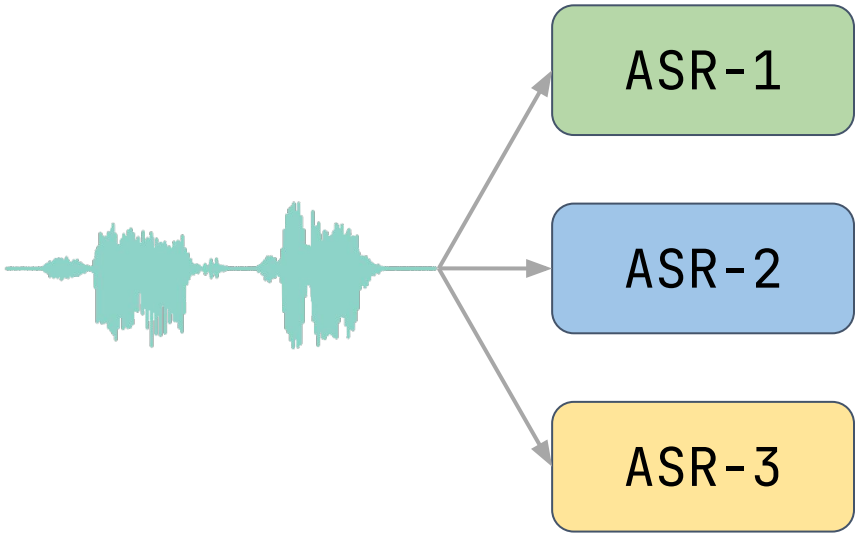


ROVER

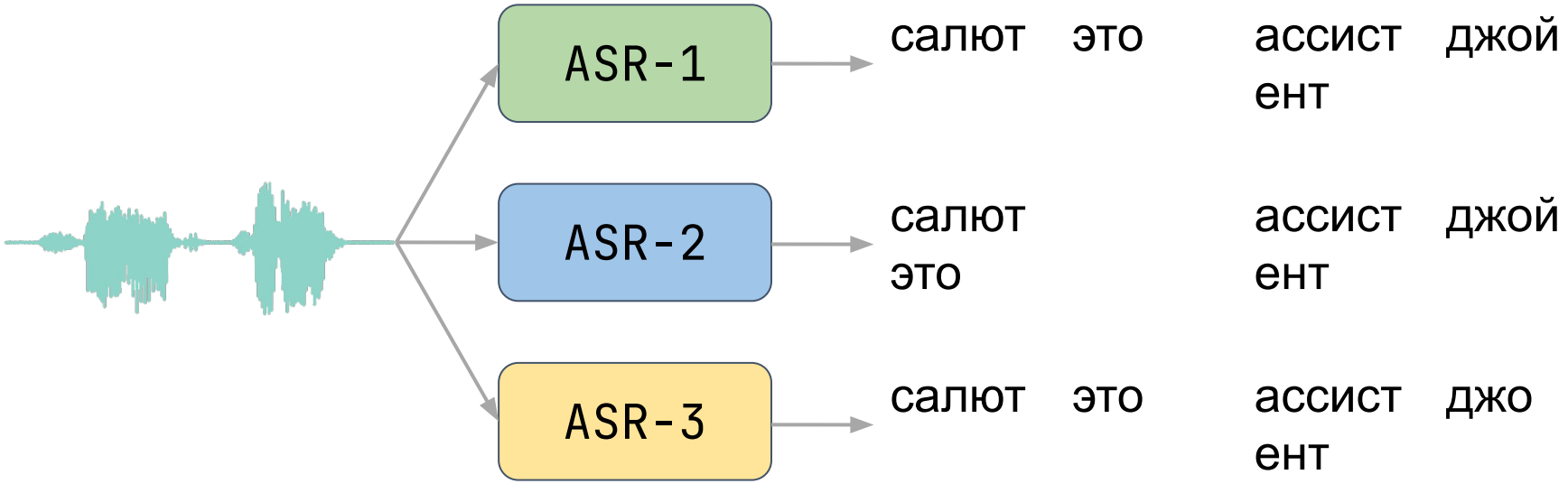


2025

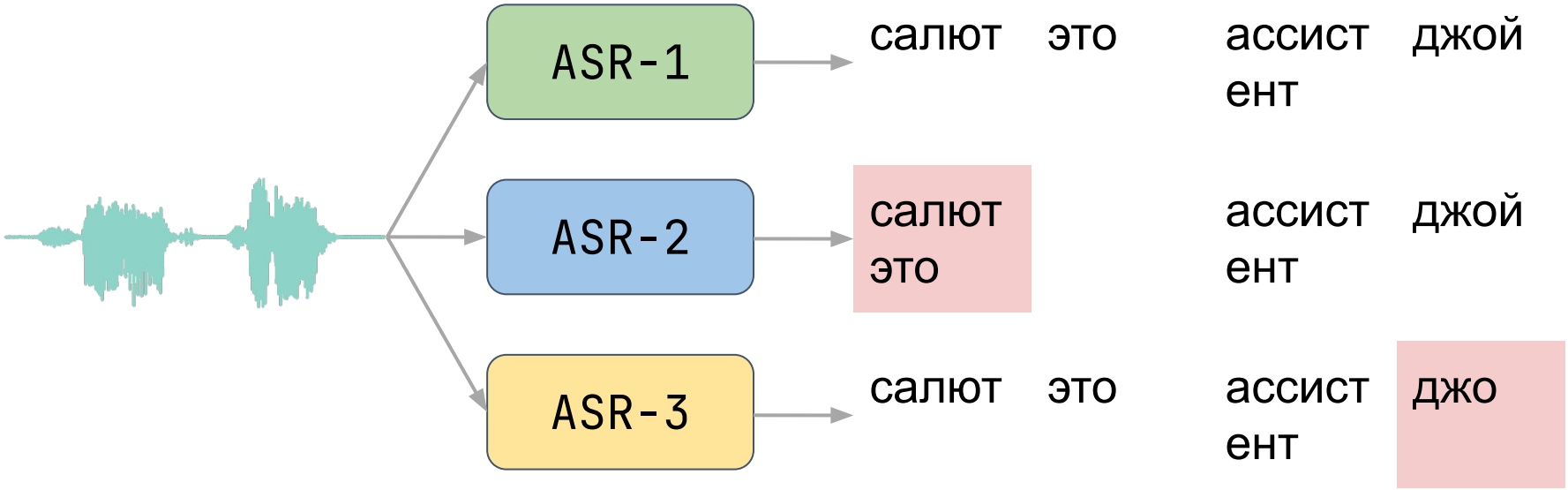
Pseudolabels: ROVER



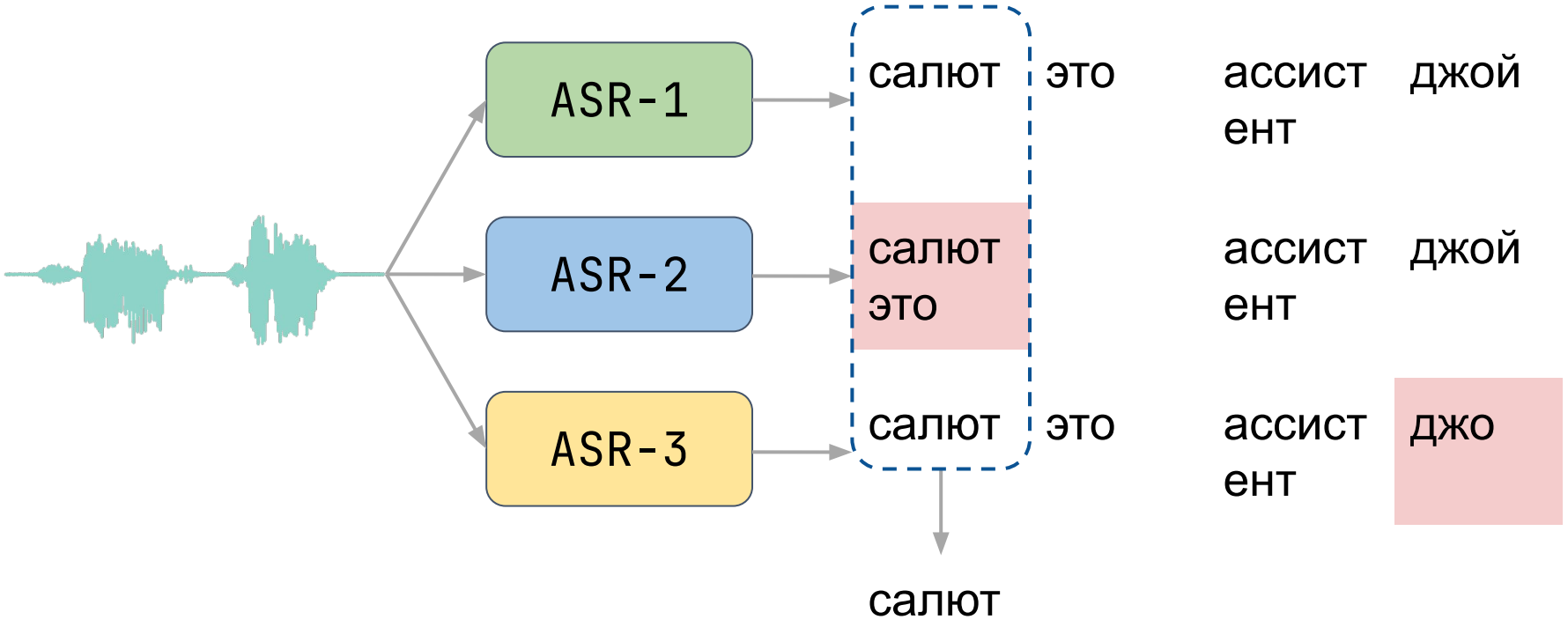
Pseudolabels: ROVER



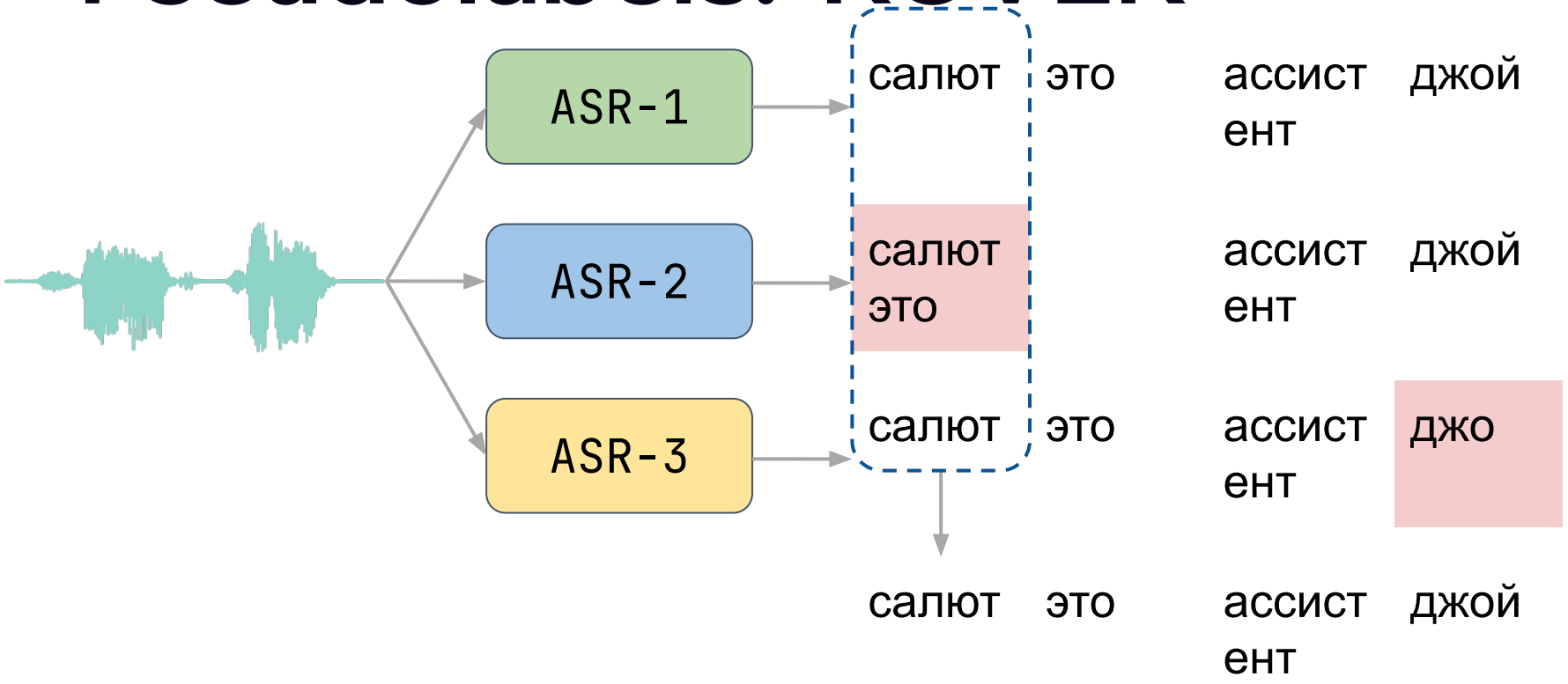
Pseudolabels: ROVER



Pseudolabels: ROVER

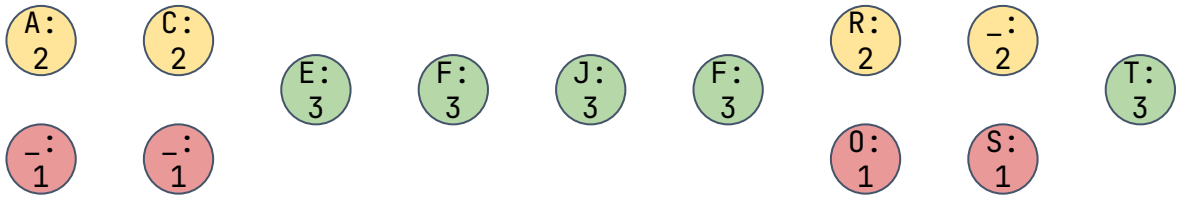


Pseudolabels: ROVER



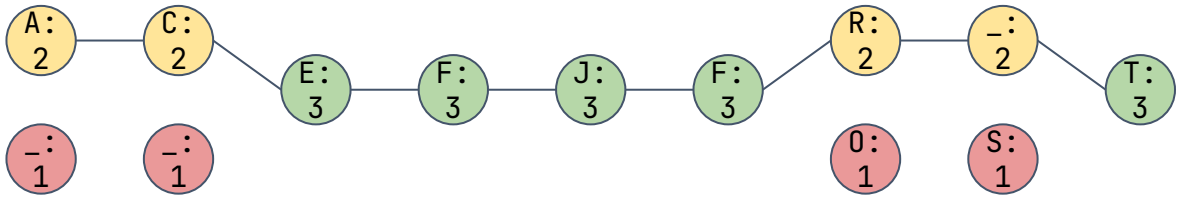
ROVER: Voting

- ACEFJFRT , ACEFJFOST, EFJFRT



ROVER: Voting

- ACEFJFRT , ACEFJFOST, EFJFRT

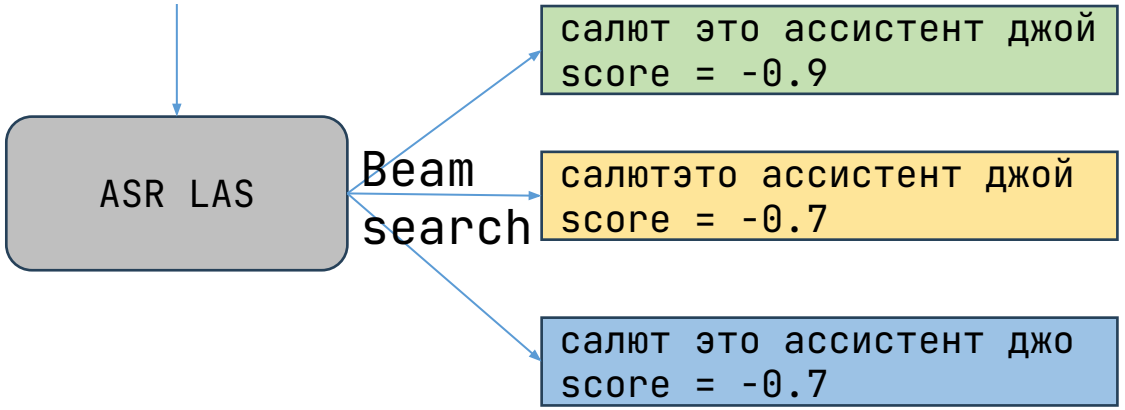


$$\text{score}(w) = \alpha \frac{N(w)}{N} + (1 - \alpha) C(w)$$

Rescoring

2025

Rescoring



2025

Rescoring

ASR LAS

Beam search

- салют это ассистент джой
  score = -0.9
- салютэто ассистент джой
  score = -0.7
- салют это ассистент джо
  score = -0.7

ASR CTC

- салют это ассистент джой
  CTC_score = -2.1
- салютэто ассистент джой
  CTC_score = -3.0
- салют это ассистент джо
  CTC_score = -2.0

Rescoring

ASR LAS

Beam search

салют это ассистент джой
score = -0.9

салютэто ассистент джой
score = -0.7

салют это ассистент джо
score = -0.7

ASR CTC

салют это ассистент джой
CTC_score = -2.1

салютэто ассистент джой
CTC_score = -3.0

салют это ассистент джо
CTC_score = -2.0

LLM

салют это ассистент джой
score = -0.9

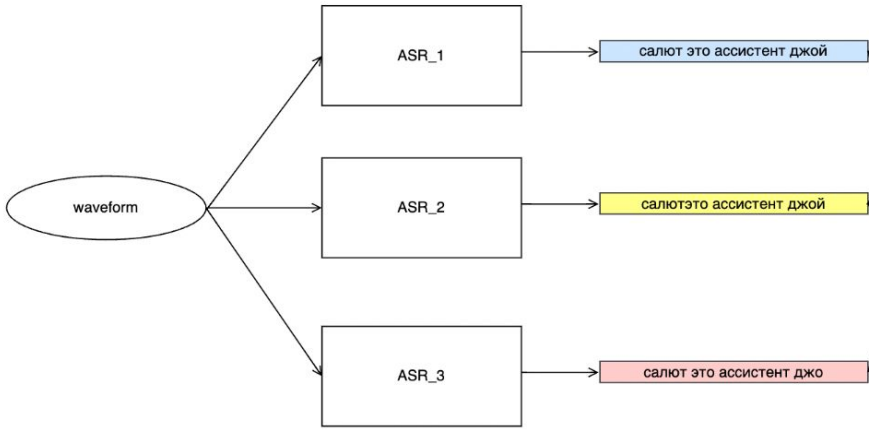
салютэто ассистент джой
score = -0.7

салют это ассистент джо
score = -0.7

2025

MBR: System Combination

$$w^* = \arg\min_{w'} \sum_{m=1}^{M} \lambda_m \sum_{w \in \mathcal{N}} \mathcal{L}(w', w) \frac{P_m(w|\mathbf{x})}{\sum_{\hat{w}} P_m(\hat{w}|\mathbf{x})}$$



2025

MBR: Minimum Bayes' Risk

$$w^* = \arg\min_{w'} \sum_{w \in \mathcal{N}} \mathcal{L}(w', w) \frac{P(w|\mathbf{x})}{\sum_{\hat{w}} P(\hat{w}|\mathbf{x})}$$

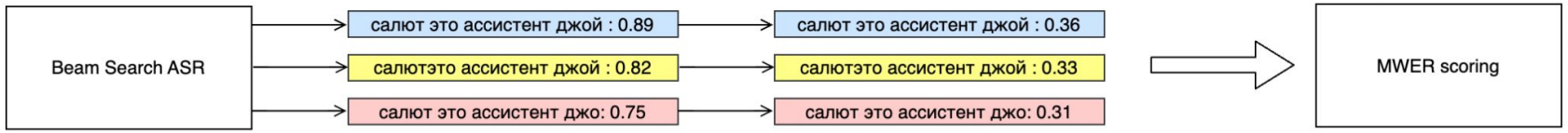
2025

MBR: Minimum Bayes' Risk

$$w^* = \arg\min_{w'} \sum_{w \in \mathcal{N}} \mathcal{L}(w', w) \frac{P(w|\mathbf{x})}{\sum_{\hat{w}} P(\hat{w}|\mathbf{x})}$$

Edit Distance Matrix

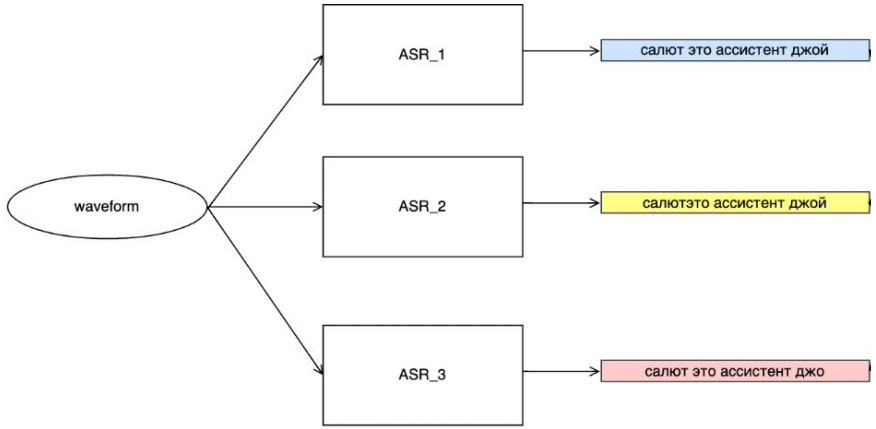
0	2	1
	0	3
		0



hypothesis	expected WER computation	expected WER
салют это ассистент джой	0 * 0.36 + 2 * 0.33 + 1 * 0.31	0.97
салютэто ассистент джой	2 * 0.36 + 0 * 0.33 + 3 * 0.31	1.65
салют это ассистент джо	1 * 0.36 + 3 * 0.33 + 0 * 0.31	1.35

MBR: System Combination

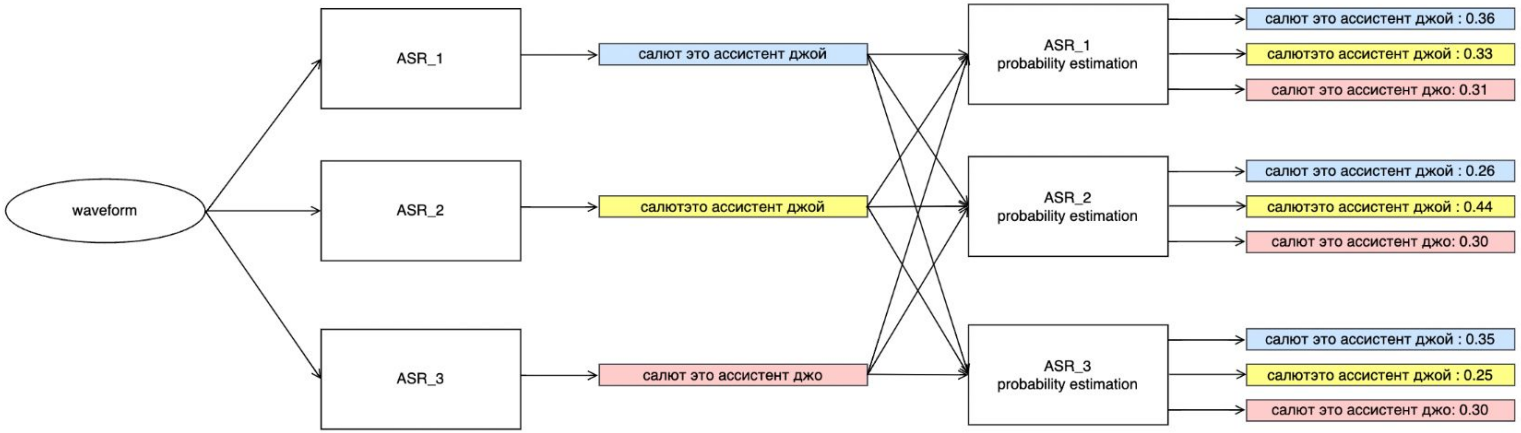
$$w^* = \arg\min_{w'} \sum_{m=1}^{M} \lambda_m \sum_{w \in \mathcal{N}} \mathcal{L}(w', w) \frac{P_m(w|\mathbf{x})}{\sum_{\hat{w}} P_m(\hat{w}|\mathbf{x})}$$



2025

MBR: System Combination

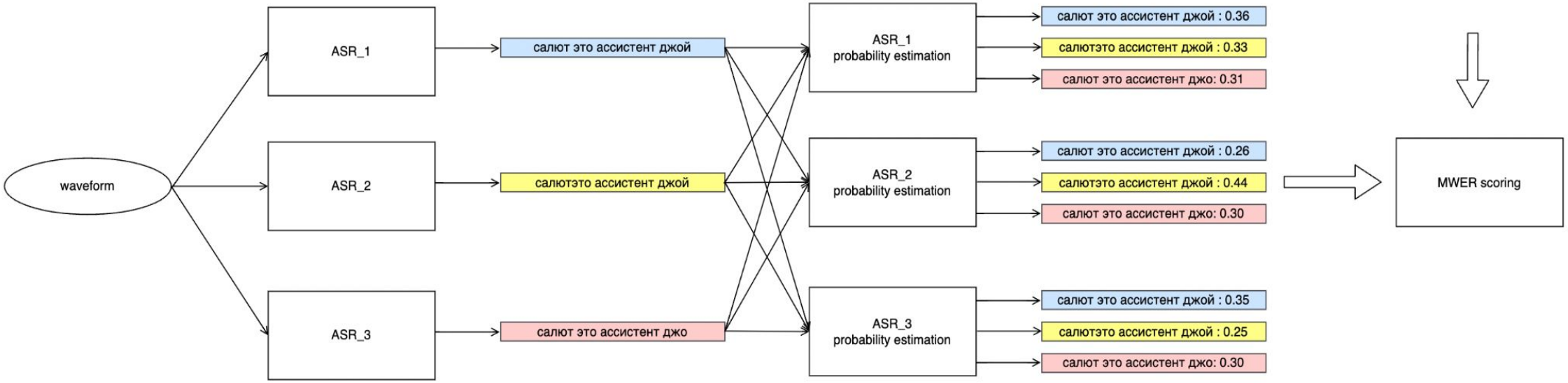
$$w^* = \arg\min_{w'} \sum_{m=1}^{M} \lambda_m \sum_{w \in \mathcal{N}} \mathcal{L}(w', w) \frac{P_m(w|\mathbf{x})}{\sum_{\hat{w}} P_m(\hat{w}|\mathbf{x})}$$

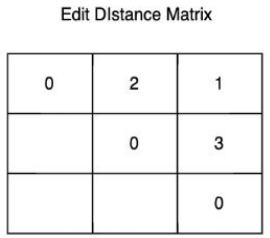


2025

MBR: System Combination

$$w^* = \arg\min_{w'} \sum_{m=1}^{M} \lambda_m \sum_{w \in \mathcal{N}} \mathcal{L}(w', w) \frac{P_m(w|\mathbf{x})}{\sum_{\hat{w}} P_m(\hat{w}|\mathbf{x})}$$

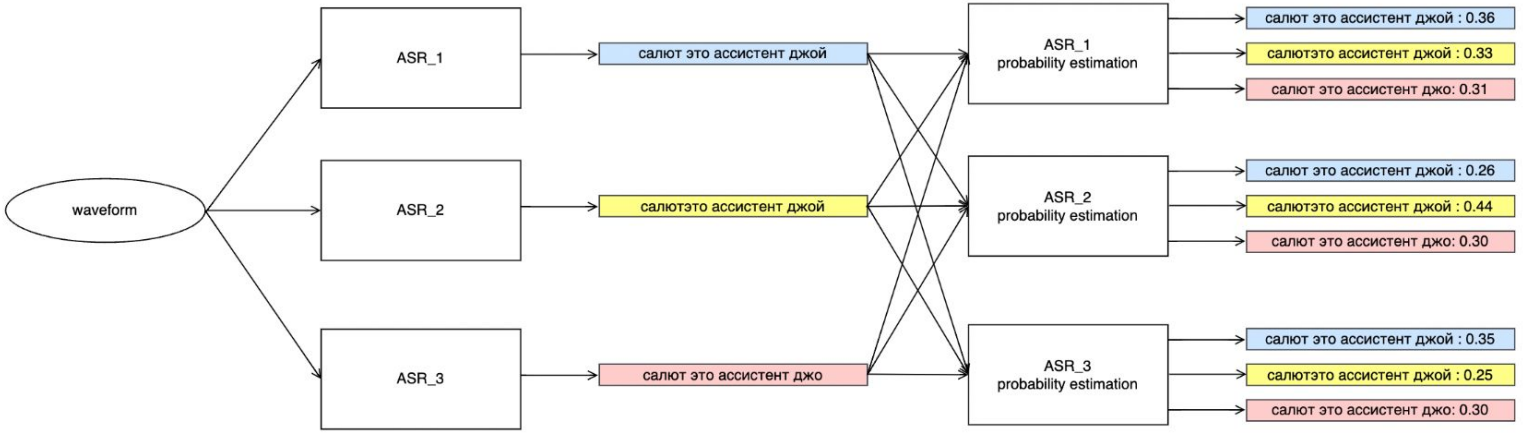




2025

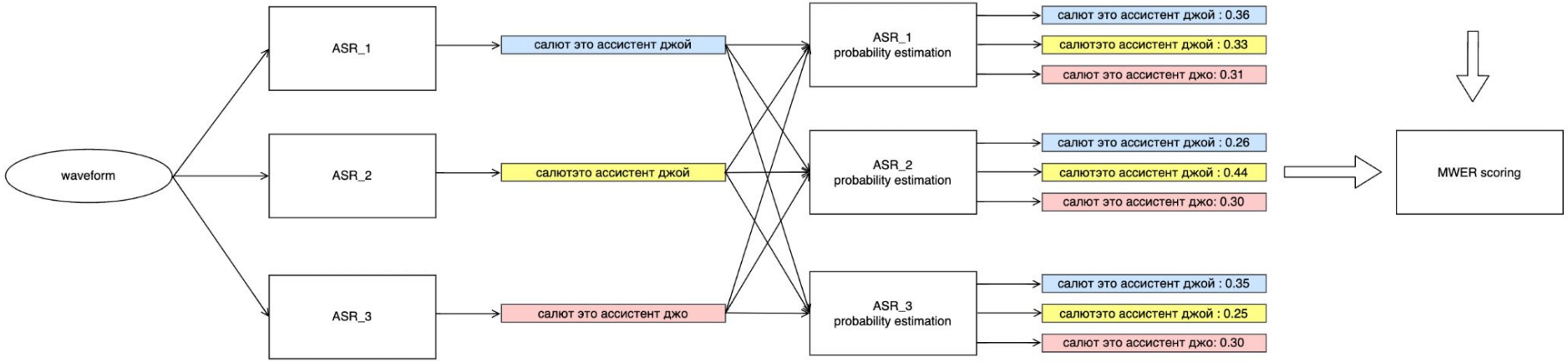
MBR: System Combination

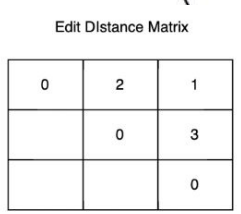
$$w^* = \arg\min_{w'} \sum_{m=1}^{M} \lambda_m \sum_{w \in \mathcal{N}} \mathcal{L}(w', w) \frac{P_m(w|\mathbf{x})}{\sum_{\hat{w}} P_m(\hat{w}|\mathbf{x})}$$



MBR: System Combination

$$w^* = \arg\min_{w'} \sum_{m=1}^{M} \lambda_m \sum_{w \in \mathcal{N}} \mathcal{L}(w', w) \frac{P_m(w|\mathbf{x})}{\sum_{\hat{w}} P_m(\hat{w}|\mathbf{x})}$$





2025

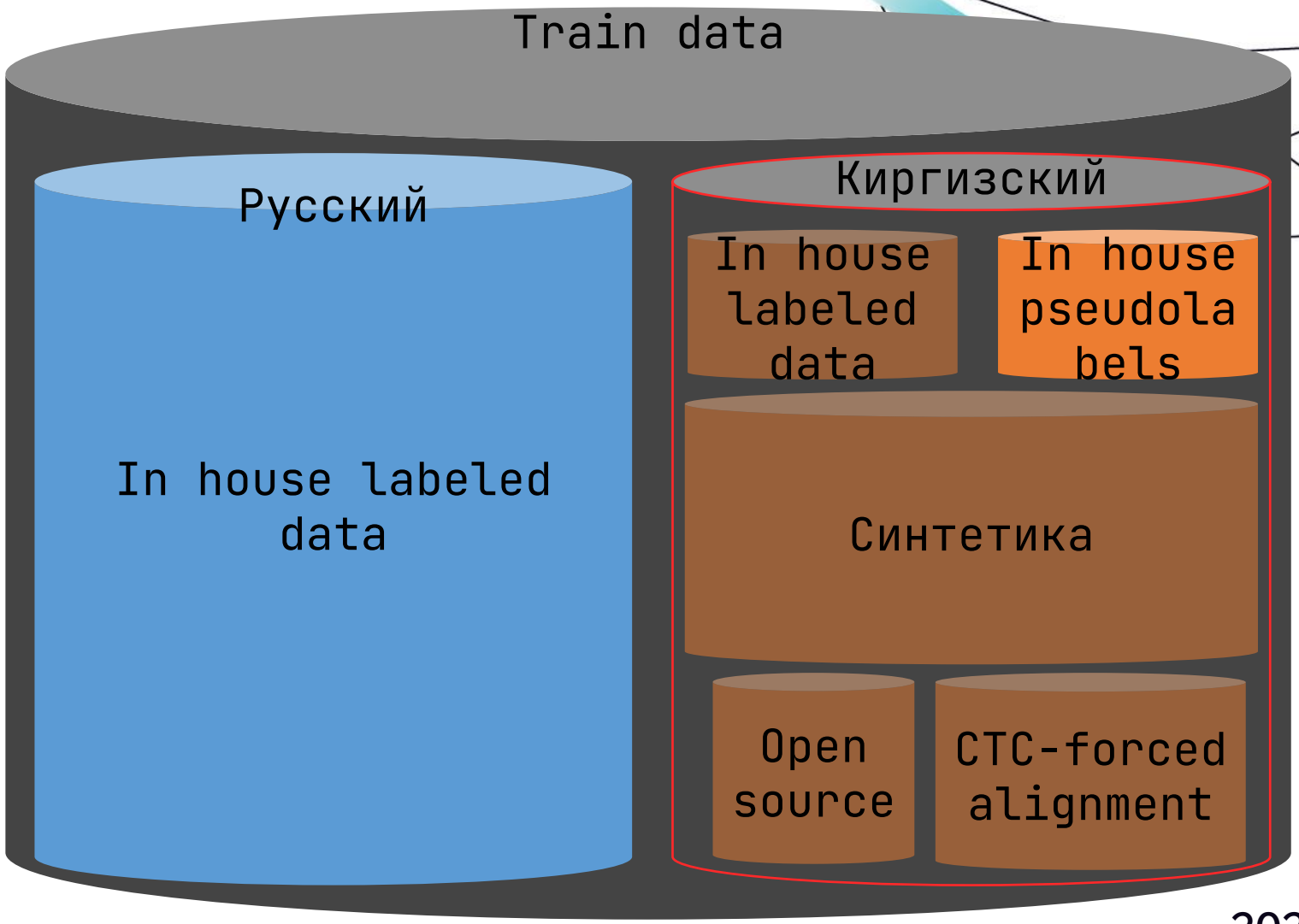
Данные

Русский (high resource)

- In house labeled data

Киргизский (low resource)

- Open source
- CTC-forced alignment
- Синтетика
- In house labeled data
- Pseudolabels



2025

Pseudolabels: результат

	common voice	fleurs	общий домен (in house)	узкий домен (in house)
baseline	41.2	32.5	71.8	86.3
baseline 2.0 (transfer learning)	20.5	16.0	46.8	71.8
+ CTC-forced alignment	15.3	12.9	48.1	69.1
+ TTS	12.2	10.3	42.1	68.8
+ In house mark	11.1	9.5	18.6	20.3
+ Pseudolabels	10.9	9.7	17.5	18.1

DLS

Спасибо за внимание!

Я

GigaDev

Кузьменко Андрей

ML-engineer, SBER

2025