ML System Design Встреча 1

**Изображение сгенерировано Kandinsky 3.0 по промптпу «Machine Learning System Design».*

Элен Теванян

МОВС

20 ноября 2024

Обо мне

купер

ML Unit Lead в операциях

НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

Препод,
магистр’18,
бакалавр’15

Доставка

Руководитель группы
развития алгоритмов эффективности

X5 RETAIL GROUP

Руководитель направления
алгоритмического анализа CVMx

Организационное про курс

Немного о скучном

- Мы все учимся
- Нет цели – завалить кого-то
- Просто встречаться не получится, придётся написать пару сочинений, чтобы я поставила оценку:

0.4*ДЗ + 0.6*Проект

- С ожиданиями вернусь на следующей неделе

Обсудить

Куда складируем материалы – договорились, что:

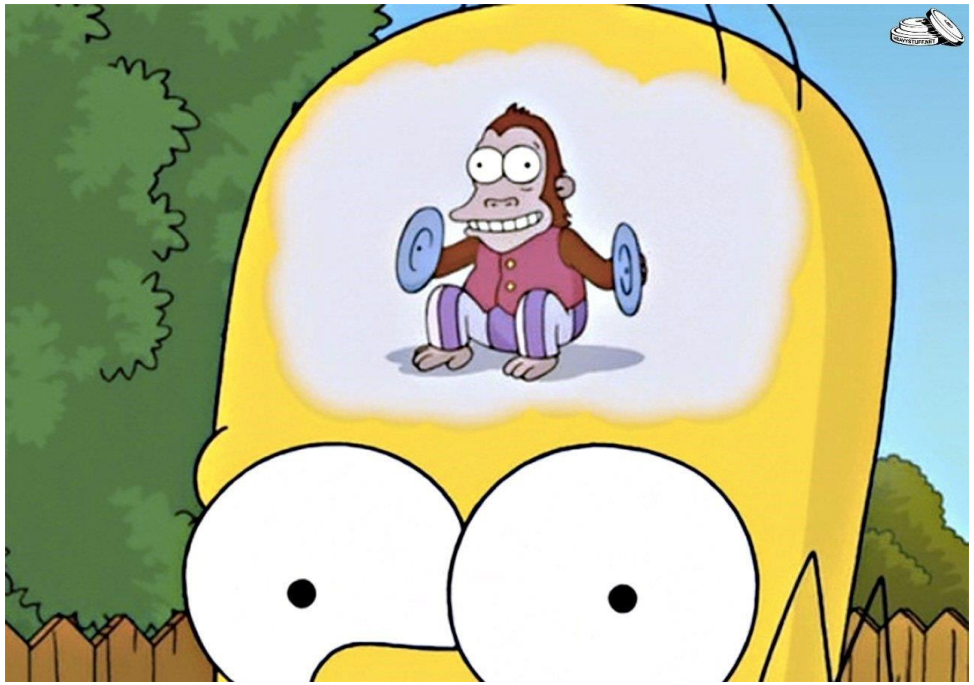
- чат
- Вики

План встреч

- 20-11-2024: Введение в дизайн систем машинного обучения
- 26-11-2024: Датасеты и инжиринг признаков
- 27-11-2024: Выбор моделей и обучение моделей
- 04-12-2024: Оценка моделей
- 11-12-2024: Диагностика проблем и мониторинги ML-систем
- 17-12-2024: Деплой ML-моделей
- 18-12-2024: Инфраструктура для машинного обучения и ML-платформы

Что такое ML System Design и с чем его едят?

Пообсуждать –
что такое ML System Design?

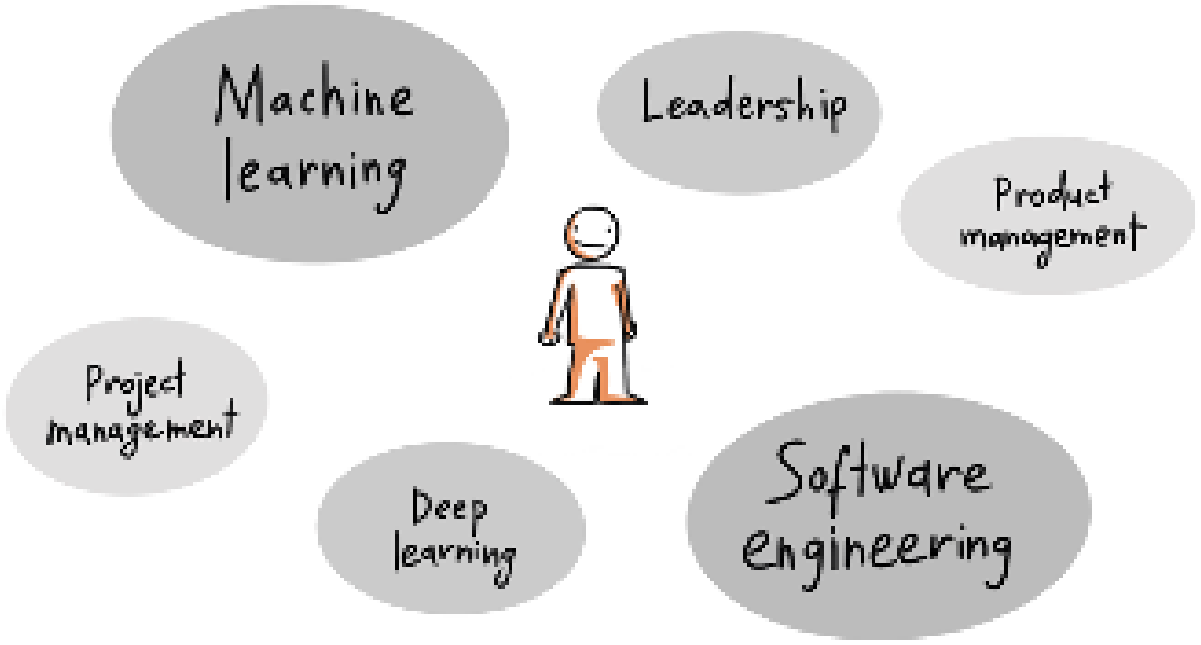


У каждого свое понимание

Что пишут в книгах

- MACHINE LEARNING SYSTEM DESIGN is a complex, multistep process of designing, implementing, and maintaining machine learning-based systems that involves a combination of techniques and skills from various fields and roles, including machine learning, software engineering, project management, product management, and leadership

[ENG] *Machine Learning System Design With end-to-end examples.* Valerii Babushkin and Arseny Kravchenko



Что пишут в книгах-2

- The system also includes the business requirements that gave birth to the ML project in the first place, the interface where users and developers interact with your system, the data stack, and the logic for developing, monitoring, and updating your models, as well as the infrastructure that enables the delivery of that logic.

[ENG] *Designing Machine Learning Systems.*
Chip Huyen (2022)

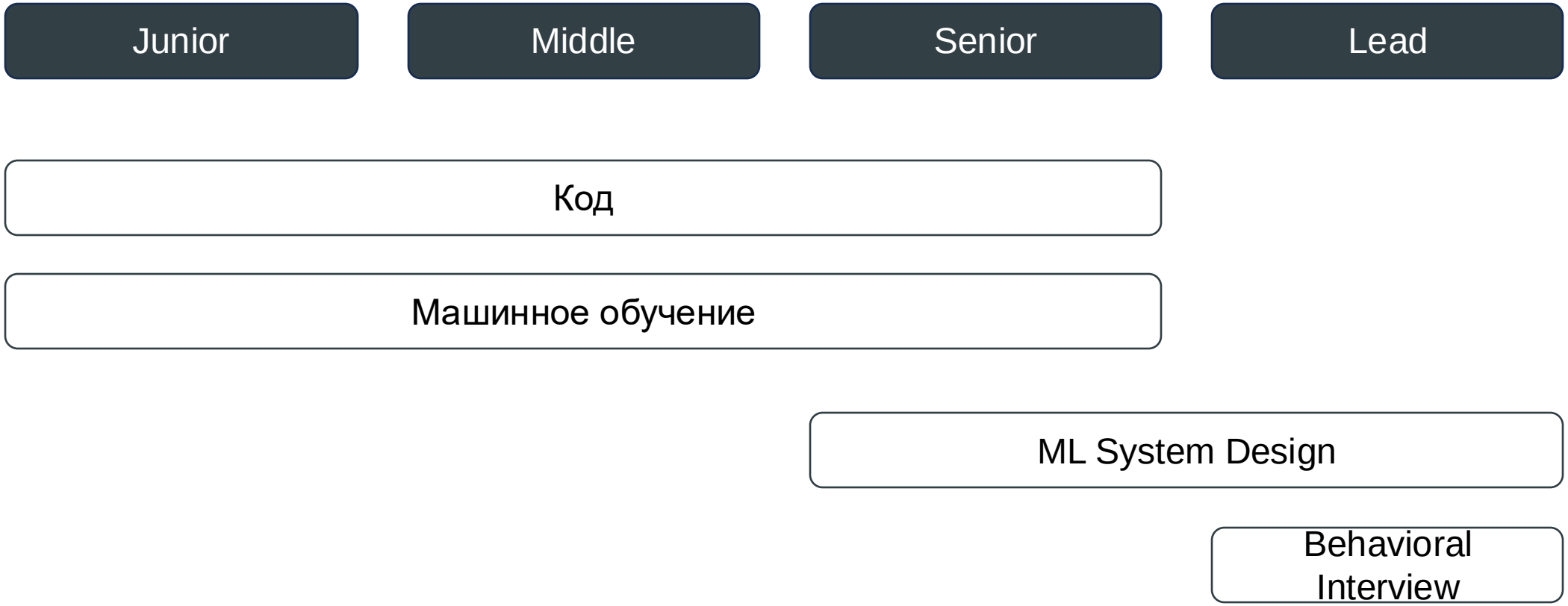
На какие вопросы хочется уметь отвечать благодаря курсу

- Модель обучена, а что дальше?
- какие компоненты нужно учесть / проработать?
- как работать с данными и фичами?
- как оценить модель онлайн и оффлайн?
- как регулярно мониторить модели и обновлять?
- …

ML System Design != ML-алгоритмы

- ML-алгоритмы – наименьшая из проблем на всём пути
- Алгоритмы (любые) должны решать существующие проблемы

MLSD появляется в жизни на зрелых уровнях карьеры



ML в исследованиях и ML в продакшне

Ожидание от работы в ML в индустрии

- Собрать данные
- Обучить модель
- Внедрить

Реальность работы в индустрии

- Договориться, какую метрику оптимизируем
- Придумать/адаптировать/выбрать лосс
- Собрать данные
- Натренировать модель
- Ужаснуться качеству, набрать новых данных, переразметить
- Обучить
- Еще раз ужаснуться качеству, добрать данных
- Обучить
- Задеплоить
- Крепко спать, пока идет эксперимент
- Проснуться в 4 утра от звонка инцидент-менеджментов, что поехали метрики, отменить раскатку
- Провести анализ ошибок, дообучить модель
- Задеплоить
- Поставить свечку
- Продакшн не упал, а бизнесовые метрики ухудшаются
- Записаться к психотерапевту
- Пересмотреть оптимизируемую метрику
- Поздравляю, все сначала

ML для исследований и в продакшне

	Исследования	**Продакшн**
Цель	Качество модели	У каждого стейкхолдера своя цель

ML для исследований и в продакшне

	Исследования	**Продакшн**
Цель	Качество модели	У каждого стейкхолдера своя цель
Данные	Статичный слепок	Постоянно изменяются

ML для исследований и в продакшне

	Исследования	**Продакшн**
Цель	Качество модели	У каждого стейкхолдера своя цель
Данные	Статичный слепок	Постоянно изменяются
Интерпретируемость	Приятный бонус	Важен

ML для исследований и в продакшне

	Исследования	**Продакшн**
Цель	Качество модели	У каждого стейкхолдера своя цель
Данные	Статичный слепок	Постоянно изменяются
Интерпретируемость	Приятный бонус	Важен
Вычислительные особенности	Быстрое обучение, высокая пропуская способность	Быстрый инференс, малая задержка

ML-разработка != Разработка

ML-разработка != Разработка

- Код и данные связаны

ML-разработка != Разработка

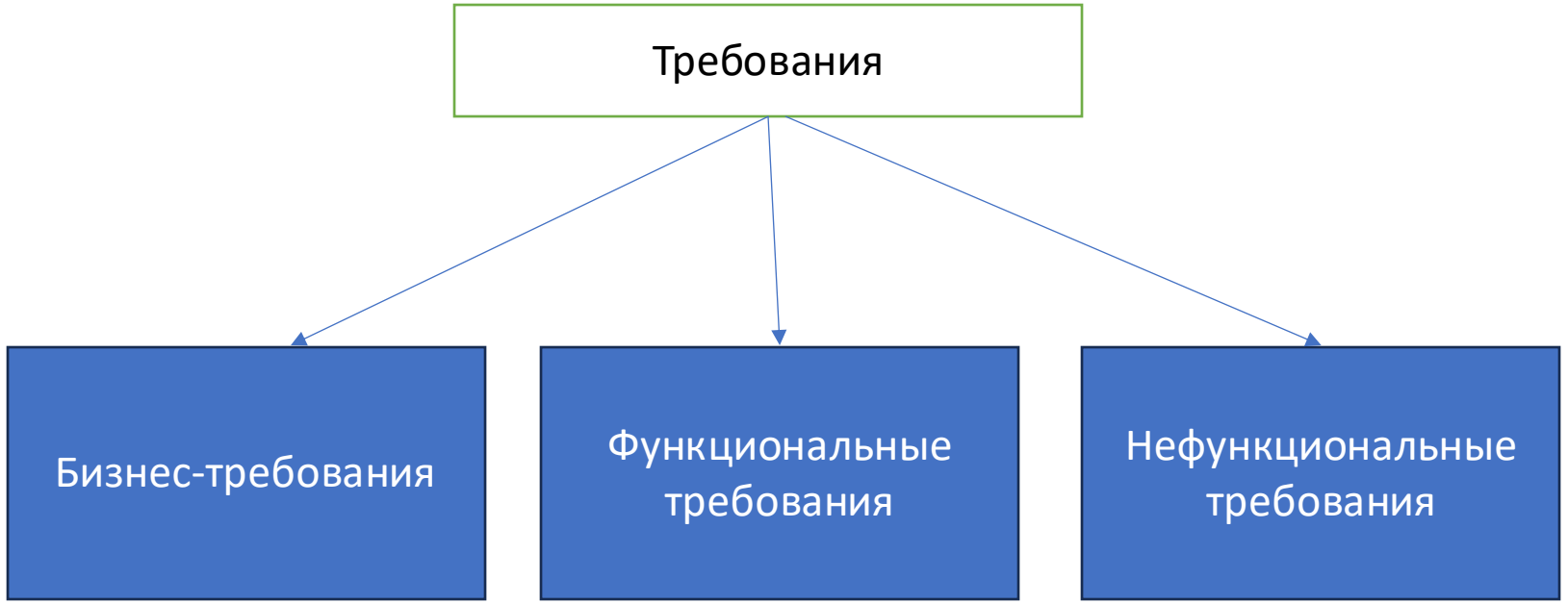
- Код и данные связаны
- Тестируется и версионируется не только код, но и данные

ML-разработка != Разработка

- Код и данные связаны
- Тестируется и версионируется не только код, но и данные
- Инженерные челленджи при работе с большими моделями

Требования к ML-системам

Требования можно разбить на три группы



В чем заключаются бизнес-проблемы, решаемые системой

Что должна делать система, её основные функции

Что ожидается от системы по качеству

Нефункциональные требования

- Realiability (надёжность)

Нефункциональные требования

- Realiability (надёжность)
- Scalability (масштабируемость)

Нефункциональные требования

- Realiability (надёжность)
- Scalability (масштабируемость)
- Maintainability (обслуживаемость)

Нефункциональные требования

- Realiability (надёжность)
- Scalability (масштабируемость)
- Maintainability (обслуживаемость)
- Adaptability (адаптируемость)

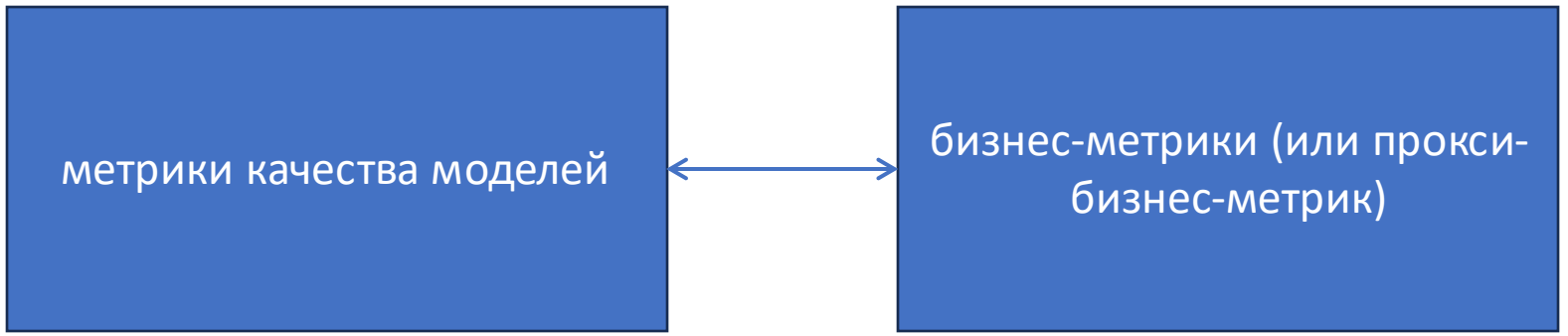
А нужен ли ML вообще?

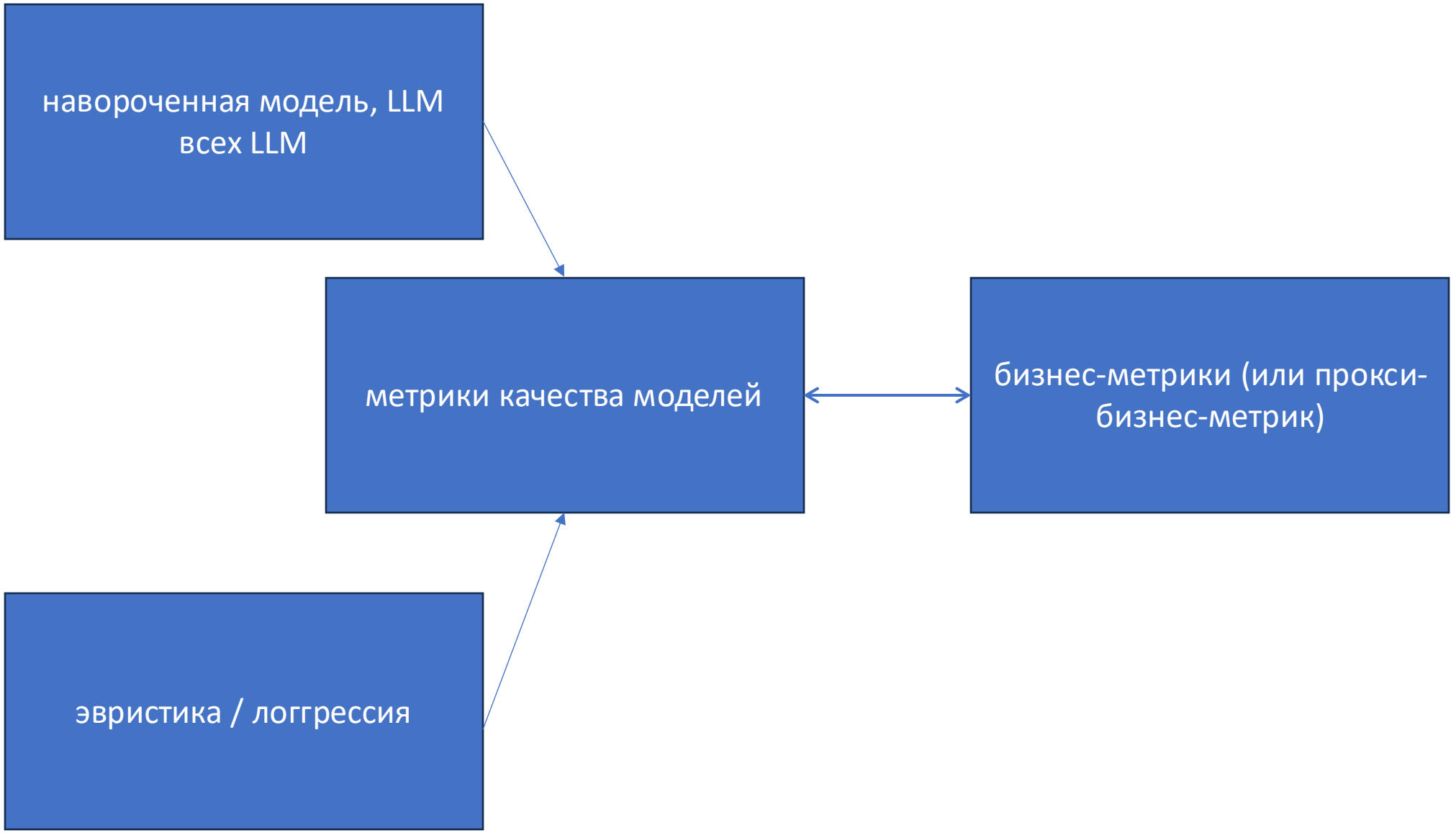
Цель бизнеса

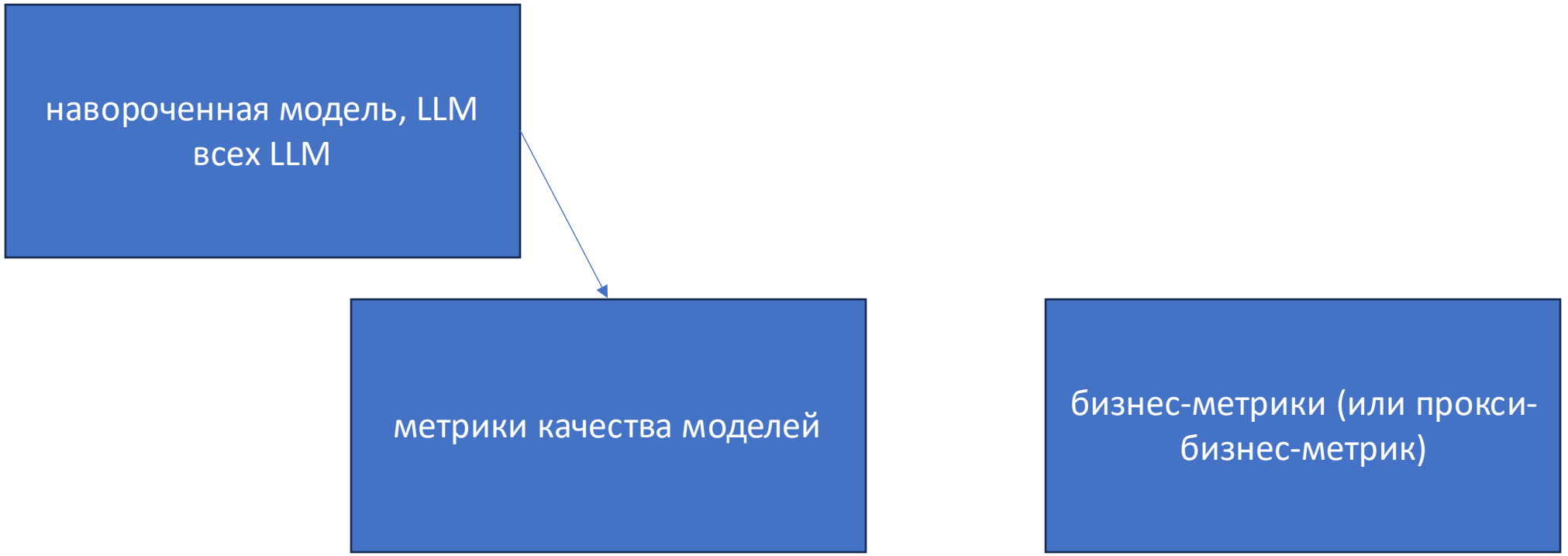
- Максимизировать прибыль
- Может принять альтернативную форму – захватить рынок, стать масштабными, чтобы выйти в прибыль и максимизировать прибыль

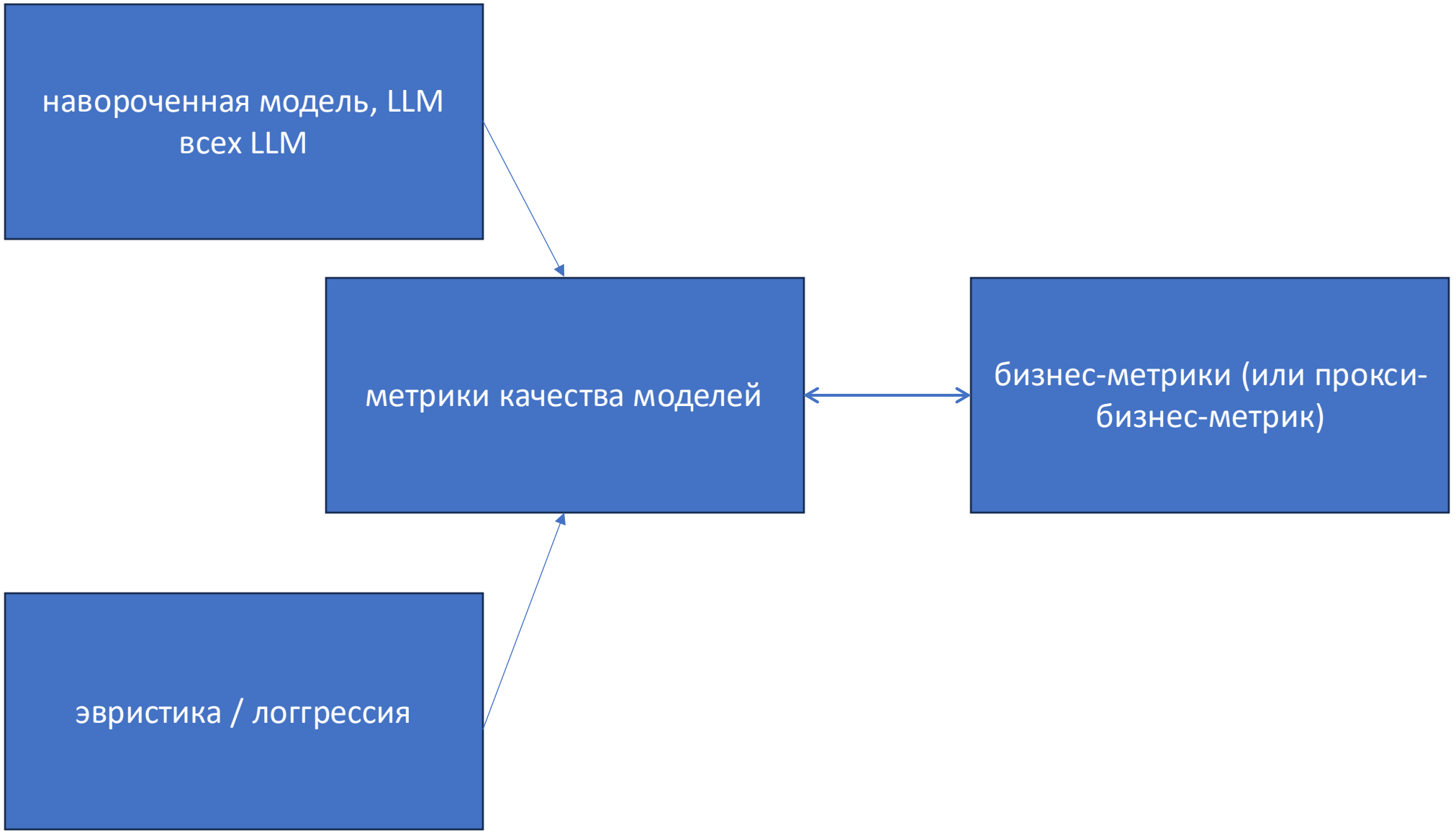
Фокус ML-команд

- Отличные метрики качества моделей
- Разумно устроенный код


метрики качества моделей
бизнес-метрики (или прокси-бизнес-метрик)

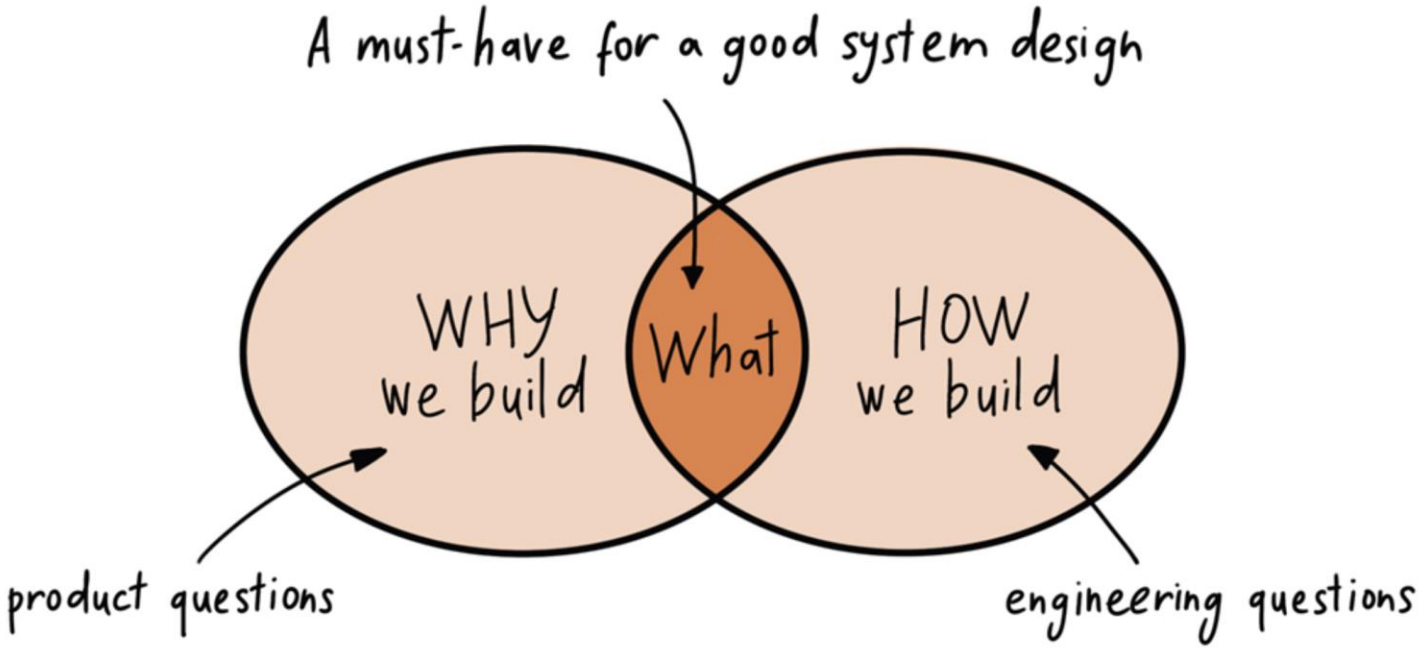

навороченная модель, LLM всех LLM
метрики качества моделей
бизнес-метрики (или прокси-бизнес-метрик)
эвристика / логгрессия


навороченная модель, LLM всех LLM
метрики качества моделей
бизнес-метрики (или прокси-бизнес-метрик)


навороченная модель, LLM всех LLM
метрики качества моделей
бизнес-метрики (или прокси-бизнес-метрик)
эвристика / логгрессия

Лучшая стратегия – стать «почемучкой»

- Почему нам нужно сделать это решение?
- Какую проблему решаем?
- Как часто появляется проблема?
- Какие альтернативы?
- Почему мы строим решения в одних ограничениях, а не других?



[ENG] *Machine Learning System Design With end-to-end examples.* Valerii Babushkin and Arseny Kravchenko

«это работа продукта»

«это работа продукта»

или все-таки ML-спецов старшего уровня

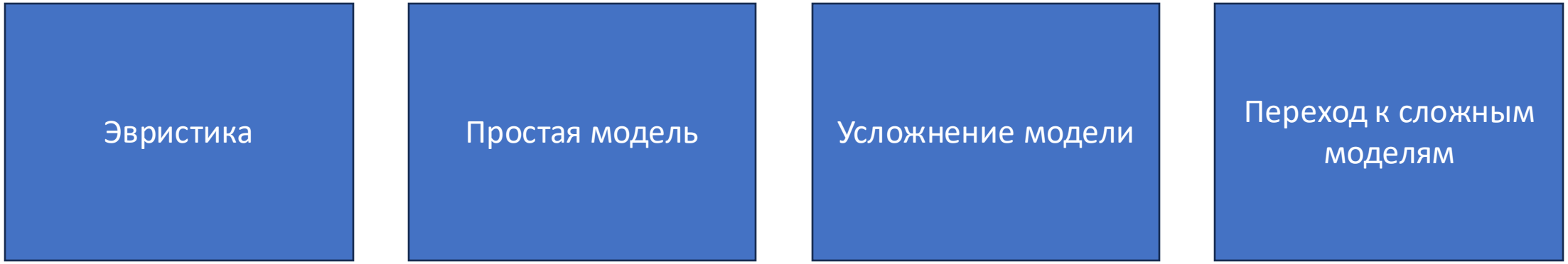
- 1 уровень – разобраться в постановке проблемы
- 2 уровень – детализировать постановку проблему почемучка-вопросами
- 3 уровень – специализированные детализированные вопросы

или все-таки ML-спецов старшего уровня

- 1 уровень – разобраться в постановке проблемы
- 2 уровень – детализировать постановку проблему почемучка-вопросами
- 3 уровень – специализированные детализированные вопросы

Что за ML-задача

4 стадии решения проблем



3 мудрых завершающих вопроса

- Что мы строим?
- Зачем мы строим?
- Как мы строим?

Всегда на связи!

@elentevanyan
https://t.me/elendatageneres