Лекция 10 — Заключительное занятие

Преподаватель: Дмитрий Меркушев

Управление ML проектами

Описание:

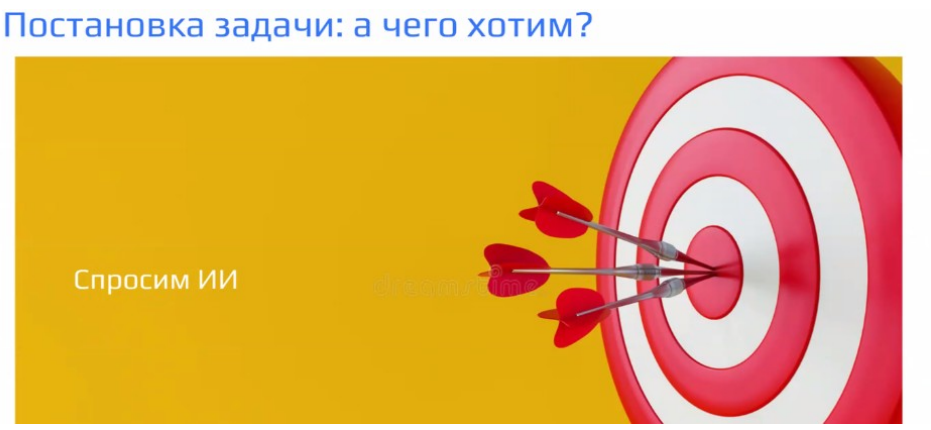
Рассмотрим типы задач, которые можно решить с помощью ML, рассмотрим, какое оборудование нужно для обучения и применения моделей.

ML проект это что-то продовое, не решается в 100% точности для ML это недостижимая история. Они уменьшают энтропию, но не убирают ее. Она заложена внутри логики.

Как обучают GPT сейчас?

1. Сбер обучают по своему(говорят)
2. Базовая модель DeepSeek берется
3. У mail-а есть GPT

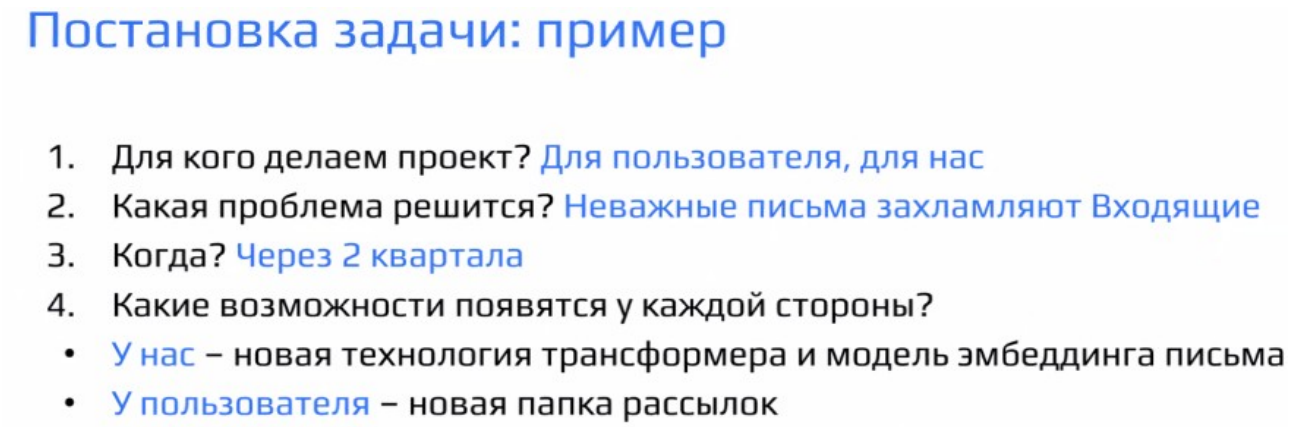
4 фазы у проекта



постановка абсолютно на всех уровнях

OKR — какие-то необычные/важные проекты оцениваются ей

ГЛАВНАЯ ОШИБКА: бросать все и решать задачи. В SMART отвечает почему / зачем и как оценивать.



Почему не надо прорабатывать просто давая решение какой-то проблемы, нужно отвечать на вопрос: хорошая ли это фича или нет.

КЛЮЧЕВЫЕ МЕТРИКИ БИЗНЕСА:

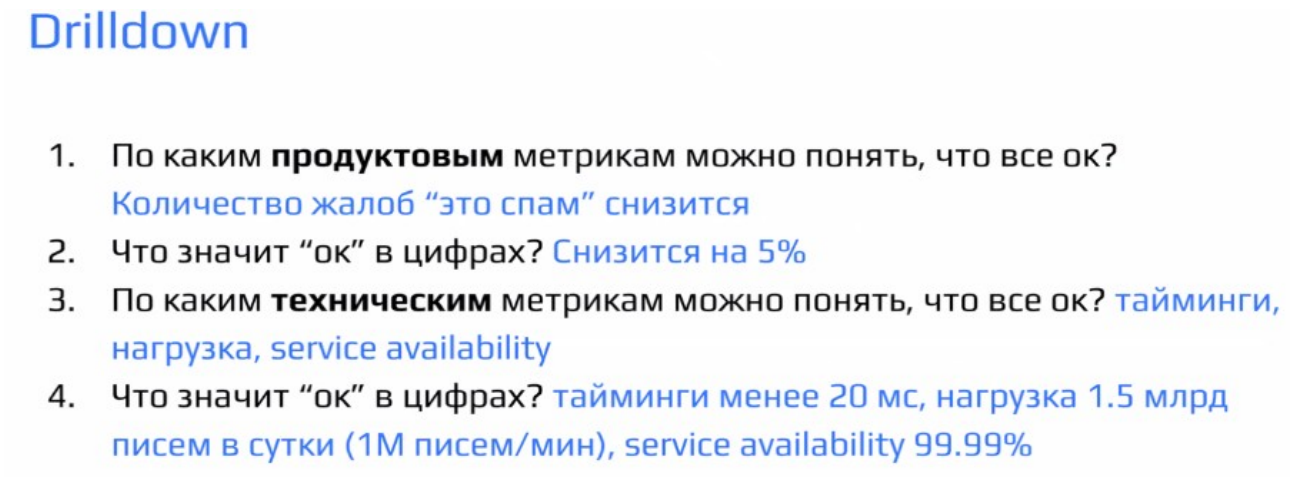
- деньги (НА рекламе и подписках). Что нужно делать? Нужно доказать ценность для пользователя и научиться материализировать все это.
- То сколько пользователь проводит времени в сервисе. (**TIME SPAN** — длительность сессии пользователя)
- Частота использования: пользователи принадлежат к ежедневному/еженедельному/ ежемесячным пользователям . Они с такой периодичностью заходят в наш сервис (ПРИМЕР: ПОЧТА)
- Почта зарабатывает на том, что при заходе на почту она выдает рекламу. Запихиваем рекламы между письмами. Донаты от пользователей/ подписки от пользователей, так же дают прибыль. Нужно подписывать сверху, что это реклама…

- Подписки: дает классный функционал (специально вшивается в подписку, чтобы кампания заработала…)

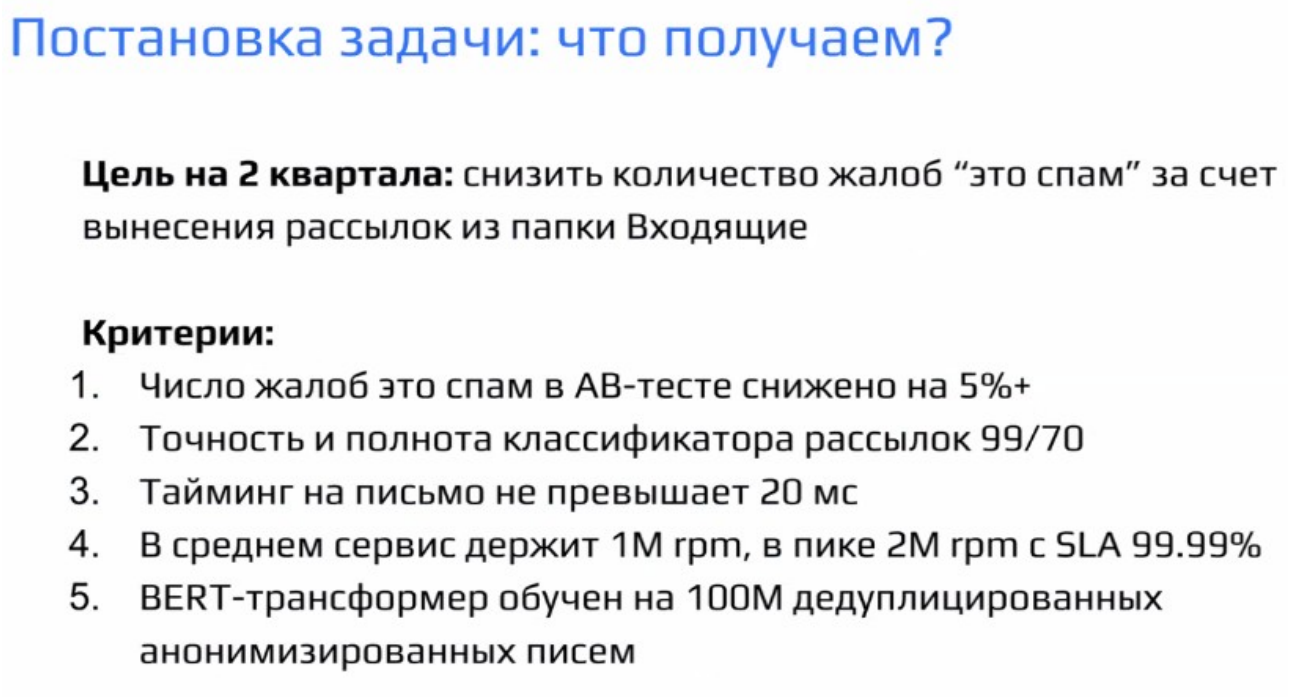
Зачем нужна категоризация писем?

- Степень удовлетворенности пользователей
- Меньше времени будет проводить в почте, т. е. пользователей будет меньше тратить время на бесполезные задачи…
- Спам - это и визуальный шум…

ЕЩЕ одна метрика — оценка пользователей, что это не спам… т. е. пользователи меньше жалуются, что это спам. Пользователь будет оставаться в нашем сервисе. Это все будет работать на ключевые метрики бизнеса.



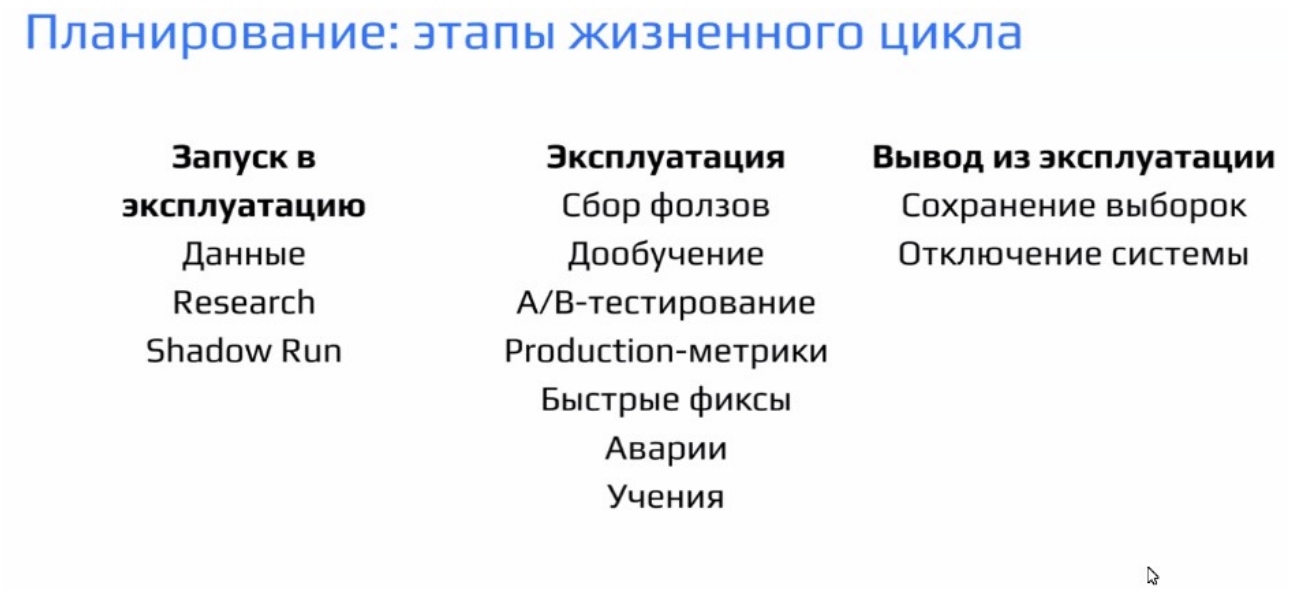
Это более точно, здесь цифры решают, но на некоторые вопросы нельзя дать четкого ответа



Анонимизация — модель будет слушать персональные данные. Нужно для обучения генеративных моделей… Иначе будет плохо… Мы должны заменять имя и данные пользователя на какие-то общее название, это улучшит модель.

Дедуплицирование -

Мы находимся в точке А и думаю о точке В



- shadow run — чтобы пользователь не понимал, что мы уже ищем данные для нашей модели.
- Выход из ресерча с моделью осуществляется — это все создание системы…
- Затем мы эксплуатируем это все и проверяем.

Фолзы — false positive
Важные метрики: Precciision (100% - не будет вообще false positive-а) / Recall

Собираем кейсы, где моделька ошиблась и дообучаем модель.

A/B тестирование — тестируем профит нашей модели.

Планирование: направления разработки

Направления:

1. Hardware
2. Back-end
3. Infrastructure
4. ML/Logics
5. Product Features
   - Back-end
   - Front-end
   - Testing/QA

Hardware — работает с железом
INFEASTRUCTURE - база
PRODUCT FEATURES- обеспечивают виденье продуктовое/ дают метрики для оценки моделей.

Все нормальные команды сходятся создать **диаграмма ГАНТА**

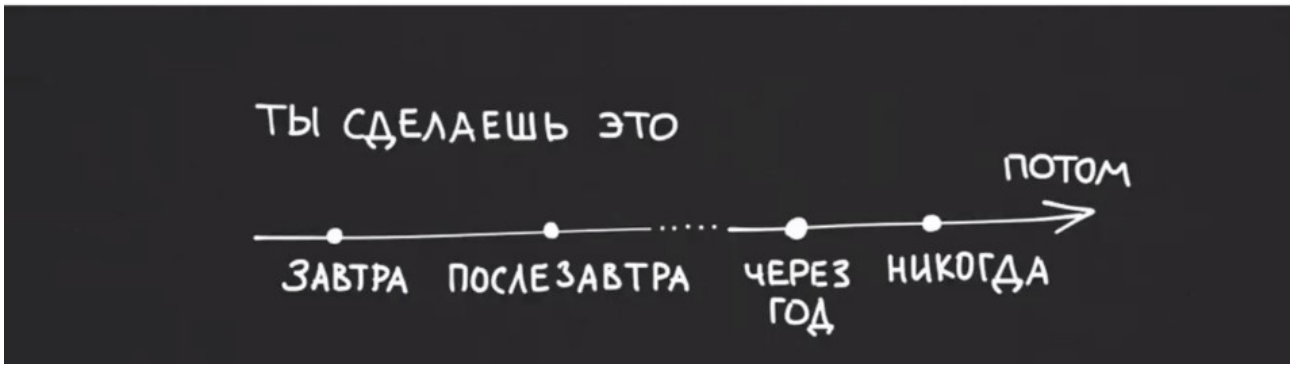
ML — это энтропийная профессия/ он не дает гарантии.

ПОСЛЕ ДИАГРАММЫ ГАНТА ПОЯВЛЯЕТСЯ МАТРИЦА:

Scope: пример

	Hardware	Backend	Infra	ML/Logics	Product Features
Shadow Run	Закупка машин	Внедрение новой архитектуры в inference-сервис	Доставка тяжелых моделей в прод	Выбрать intrinsic-метрики	Обучить 1 продуктовый клф, 1 для антиспама
Эксплуатация					
Аварии	Поведение при аварии в датацентре	Поведение при отключении inference-сервиса	Поведение при накатывании кривой модели	Превентивные и реактивные меры при утере выборок	
Вывод из эксплуатации					
Сохранение данных			Сохранение raw data	Сохранение распределений клф на бою	

Планирование: milestones



ЦЕЛЬ нужно декомпазировать:

Постановка задачи: что получаем?

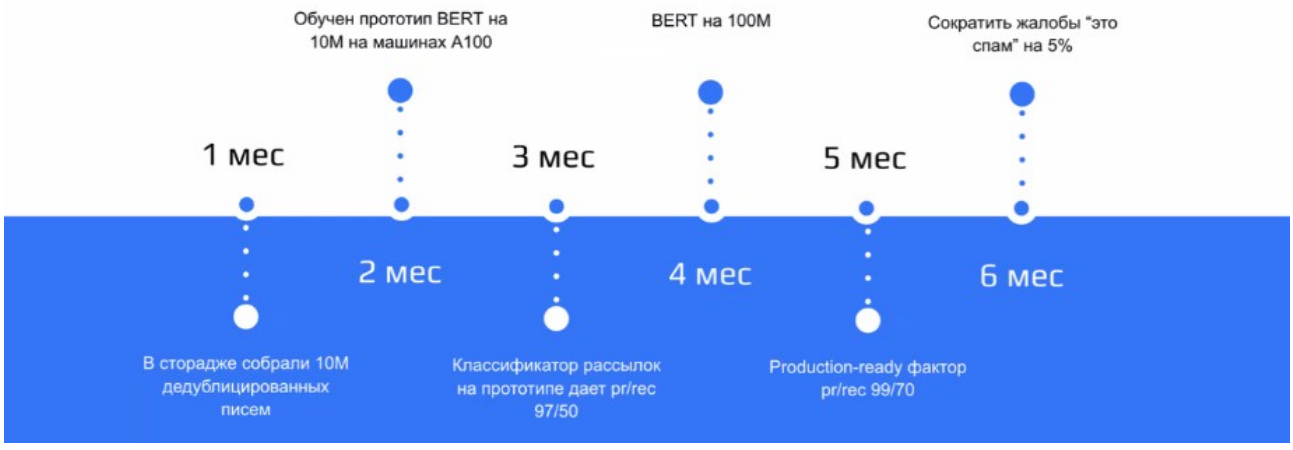
Цель на 2 квартала: снизить количество жалоб “это спам” за счет вынесения рассылок из папки Входящие

Цель на 1 квартал: получить точный прототип классификатора рассылок на BERT-трансформере промежуточного размера

Критерии:

1. BERT-трансформер обучен на 10M (100M) дедуплицированных анонимизированных писем
2. Точность и полнота классификатора рассылок 97/50 (99/70)

Milestones: пример



Кампания должна знать сколько она будет зарабатывать через какой-то промежуток времени, чтобы планировать заранее.

РАБОТА С РИСКАМИ — это ключевое в IT

Планирование: How To?

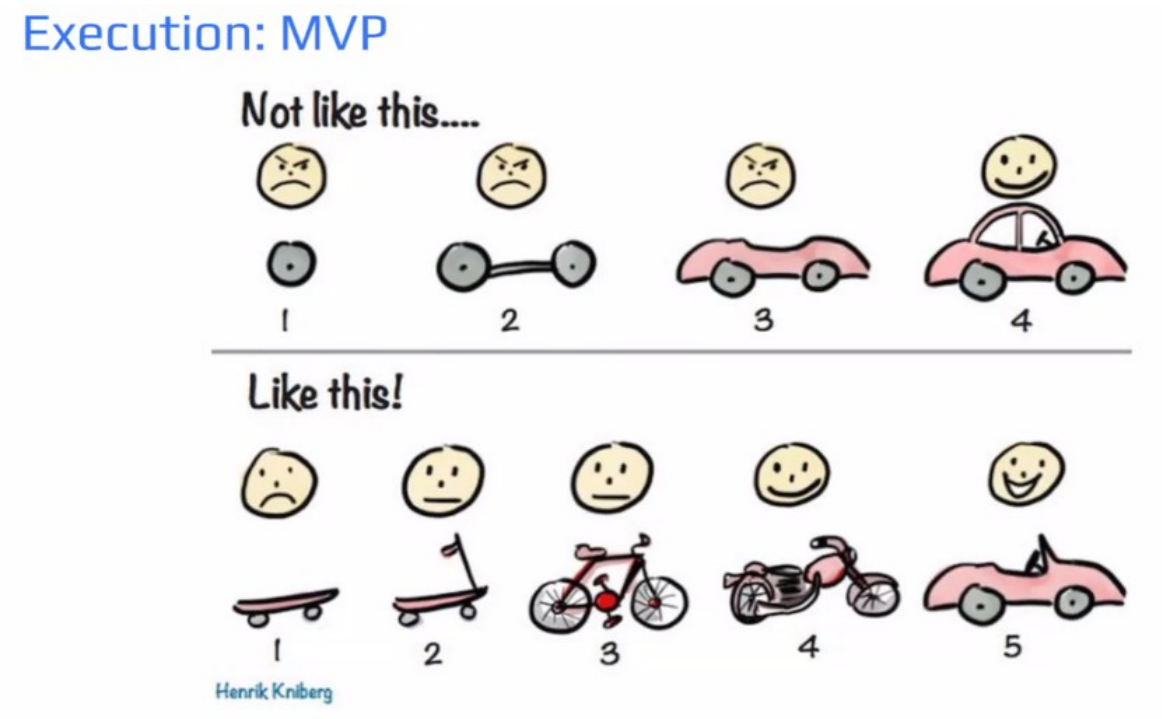
Шаги:

1. Берем ближайший майлстоун
2. Из таблицы берем задачи, необходимые для достижения майлстоуна (с учетом приоритетов)
3. Если до майлстоуна 6 недель, то прикидываем объем работы на 3-4 недели, которые может сделать команда, если все складывается идеально (все идеально не сложится)
4. Если не хватает ресурсов, то либо упрощаем задачу, либо ищем дополнительные руки

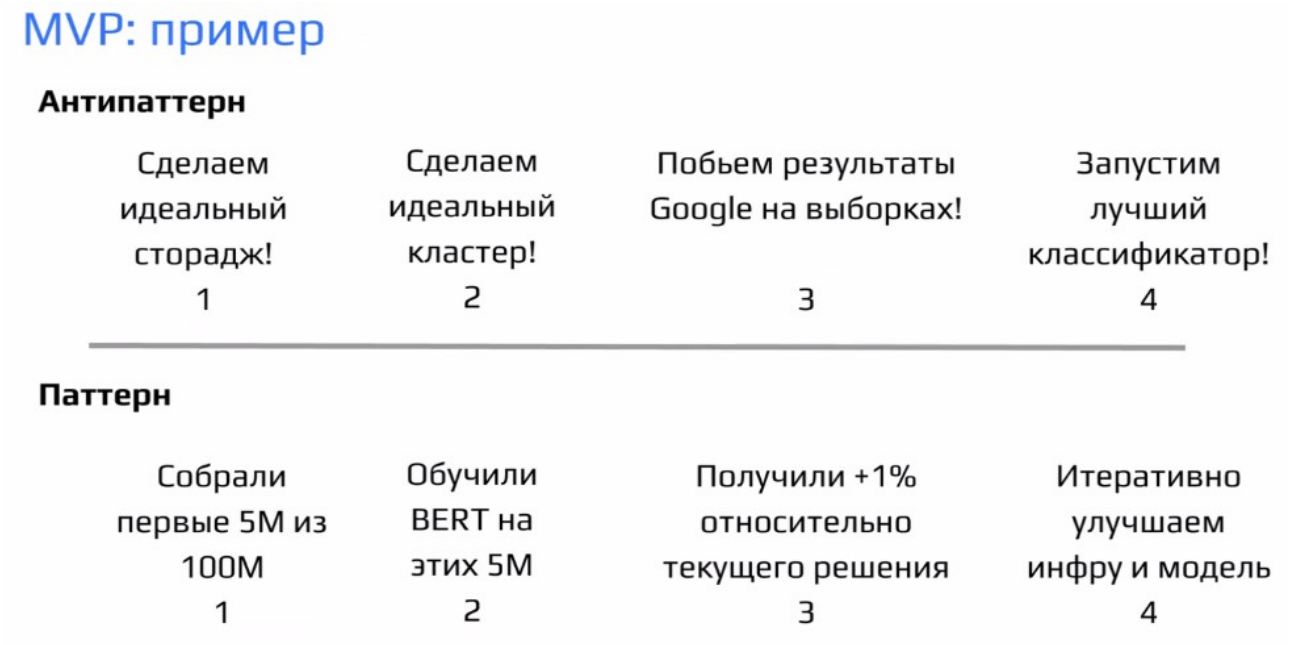
В проекте никогда не будет такое, что мы все в сроки комитем… Нужно всегда писсемизировать результат…

Лучше упростить задачу, но сделать часть задачи для того, чтобы мы закончили в строки (4 пункт - это важно) ИЛИ бери дополнительные руки...

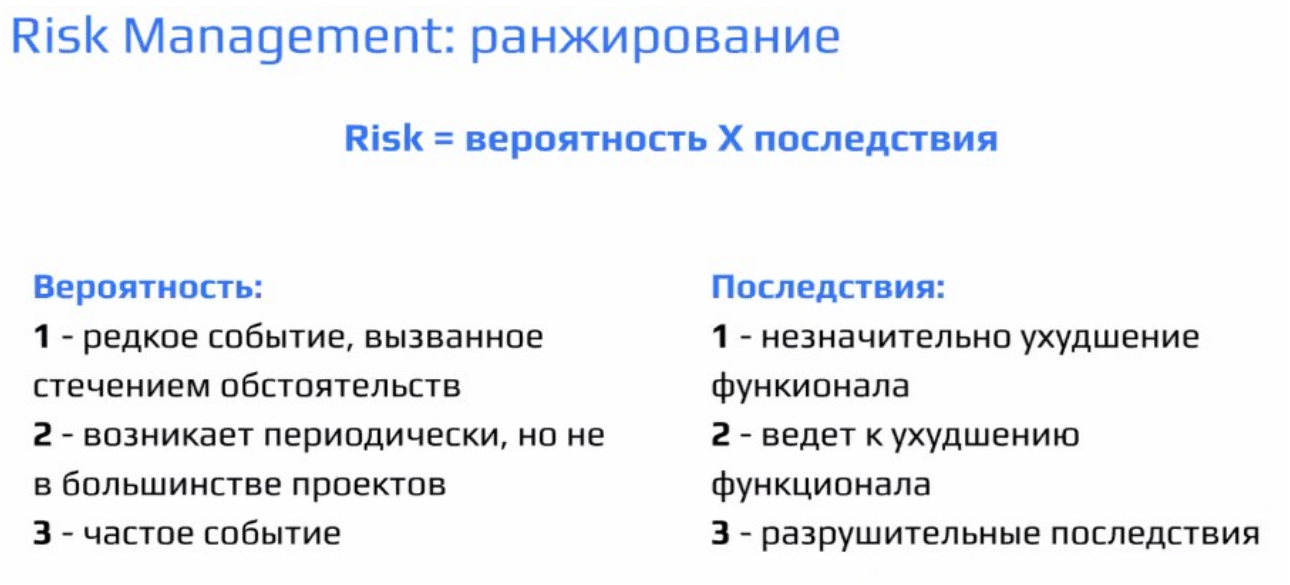
Execution: MVP



Пользователю, может только часть нашего функционала нужно. И мы так узнаем, когда можно закончить проект.



Не нужно — думать, что мы сделаем лучшее что-то…



Риск — совокупность вероятностей наступления события. (БЫВАЮТ ДВУХ ТИПОВ описано сверху)

Как понять, какие риски могут быть?

Задача: закупить сервера с GPU для обучения BERT

1. Что **должно произойти**, чтобы все было ок?
   a. Сервера должны быть доставлены и установлены в датацентре до середины проекта
   b. В стойке должно хватать питания для 3 серверов A100 с 8 GPU на борту
2. Чего **не должно произойти**, чтобы все было ок?
   a. Сервера не должны ронять во время монтажа

Список возможных рисков

Задача: закупить машины с GPU для обучения BERT

1. Машины могут быть доставлены и установлены в датацентре с опозданием (позже середины проекта)
2. В стойке не хватит питания для 3 машин A100 с 8 GPU на борту
3. Машины могут уронить во время монтажа

Risk Management: ранжирование

	Вероятность	Последствия	Rank
Машины могут быть доставлены и установлены в датацентре с опозданием (после середины проекта)	3	3	9
В стойке не хватит питания для 3 машин A100 с 8 GPU на борту	2	2	4
Машины могут уронить во время монтажа	1	3	3

Risk Management: классификация

- **Rank 1-3:** либо принимаем, либо справляемся с последствиями
- **Rank 4-6:** предотвращаем, устраняем последствия
- **Rank 7-9:** предотвращаем, проводим красные линии, план Б, устраняем последствия

Risk Management: пример

Риск: Машины могут уронить во время монтажа

Rank: 3

Шаги:

1. **Предотвращение:** заказываем одну лишнюю машину, если позволяет бюджет
2. **Митигация:** договариваемся с дружественной командой, что может быть придем за 1 машиной

Documentations

1. Репозиторий
2. Куда пишутся логи?
3. Типичные проблемы и как их решать?
4. Какие ручки в конфигах?

Alerts

1. Метрики живучести
2. Счетчики
3. Продуктовые метрики
4. Автоматические уведомления
5. Дежурство: кто за всем этим следит?

Retrospective

1. Что было хорошо?
2. Что можно улучшить?
3. Как улучшим?
4. План и ответственные