Lean
DataScience

Управление ML продуктами в LeanDS

Асхат Уразбаев

Асхат Уразбаев

- Основатель ScrumTrek
- Agile Coach
- Основатель LeanDS

Откуда хайп вокруг данных и ML?

- Данные полны инсайтов
- Эксперименты дешевле проводить на данных
- Принимаются лучшие решения
- Лучшее понимание клиентов
- …
- ***Компании исчерпали возможности конкурировать в других областях***

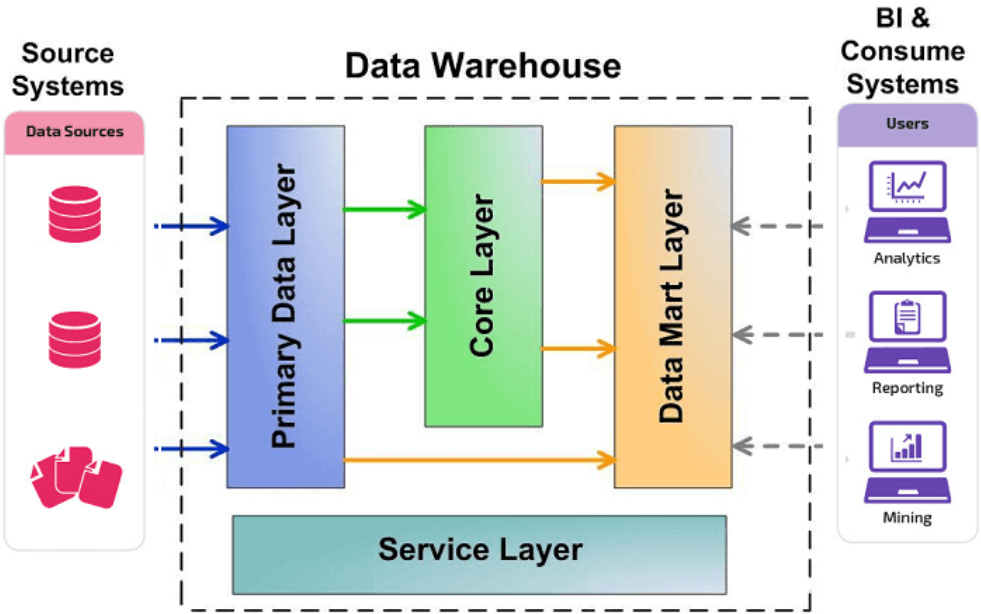
Этапы зрелости использования данных в организации

0. Data Free	1. Data Aware	2. Data Silos	3. Data Integrated	4. Optimizing
Для принятия решений менеджеры используют опыт и “чуйку”. Данным не доверяют из-за их турбулентности	Принятие долгосрочных решений: выставления KPI и квот продажников, планов производственным отделам и т.д.	Экспериментальный подход к работе с данными. Используются локально доступные данные внутри подразделений. Чаще всего это близкие к клиенту подразделения: маркетинг, продажи	Данные становятся важным активом. Подразделения начинают использовать данные, собранные по всей организации. Организация накапливает экспертизу для работы с данными.	Организация оптимизирует инфраструктуру и методологию работы с данными. Работа с данными децентрализуется.

«Традиционный» Data Warehouse

Особенности

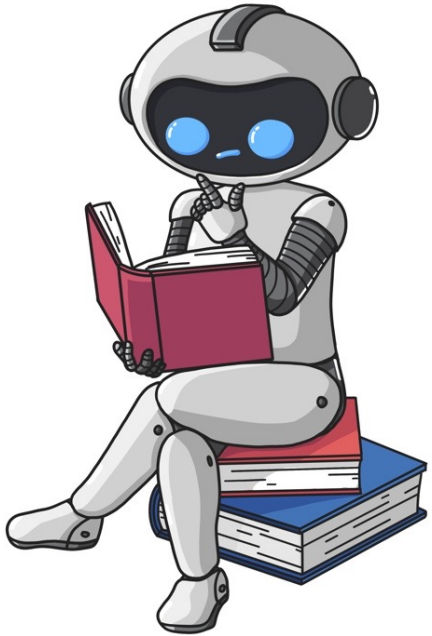
- Реляционная модель (таблички в базе данных)
- Регулярная выгрузка данных
- Фокус на отчетах для бизнеса



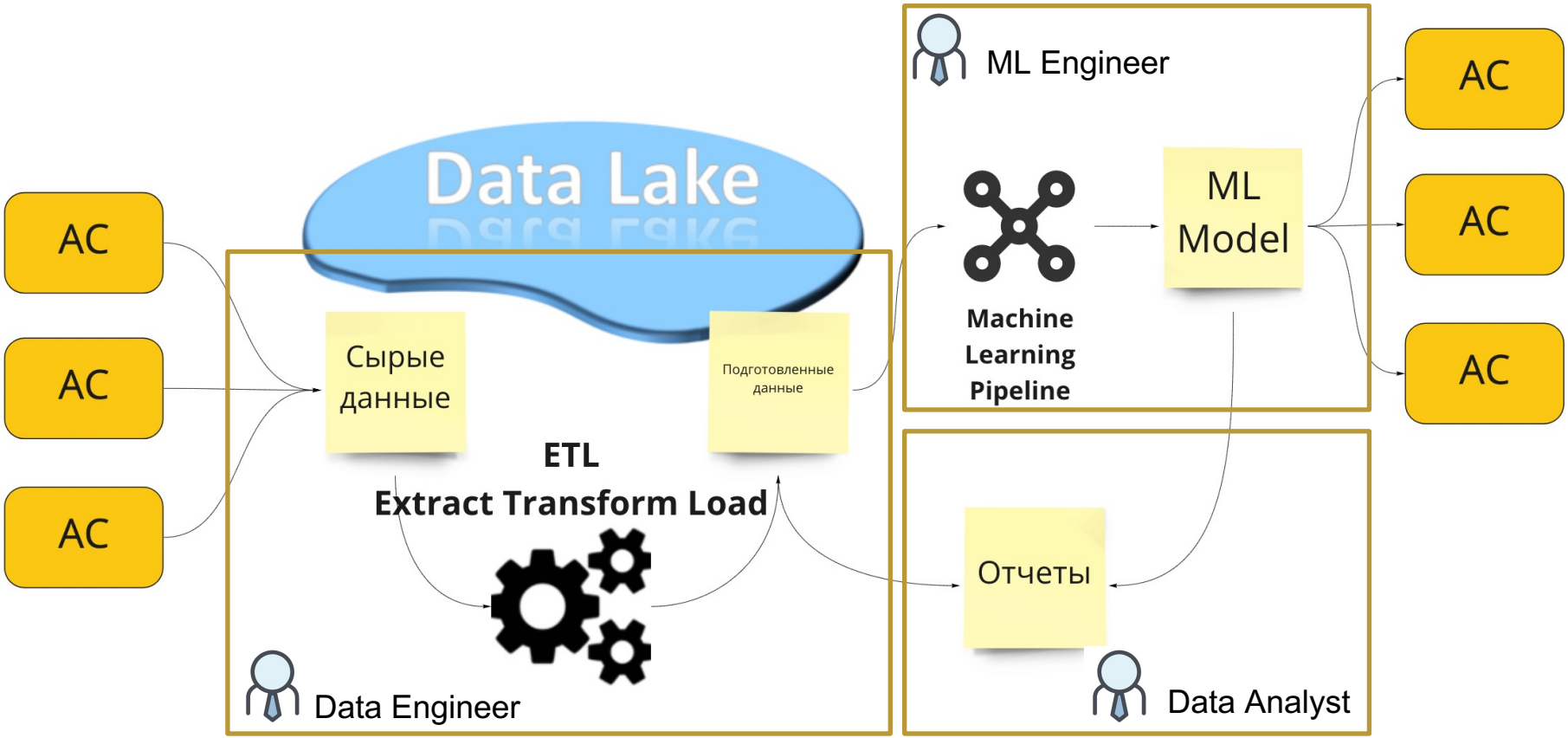
https://www.bigdataschool.ru/blog/lsa-data-warehouse-architecture.html

Приходит Machine Learning and Data Science

- **ML алгоритмы**
- Множество форматов данных (+картинки, видео, звук, неструктурированный текст, логи…)
- Real Time Data Processing
- Много-много-много данных!
- **Данные нельзя терять**
- …



Новые профессии



Примеры проблем с точки зрения бизнеса

- Новые вызовы: диджитализация маркетинга, появление новых (омни)каналов сбыта, масштабирование бизнеса, потребность в машинном обучении
- Взрывной рост потребности в отчетности со стороны новых функций бизнеса
- Традиционное хранилище данных не справляется с объемом работ, предъявляемых бизнесом
- Появление новых локальных хранилищ данных в маркетинге, производстве, логистике и машинном обучении
- Появляется потребность в переиспользовании данных между подразделениями

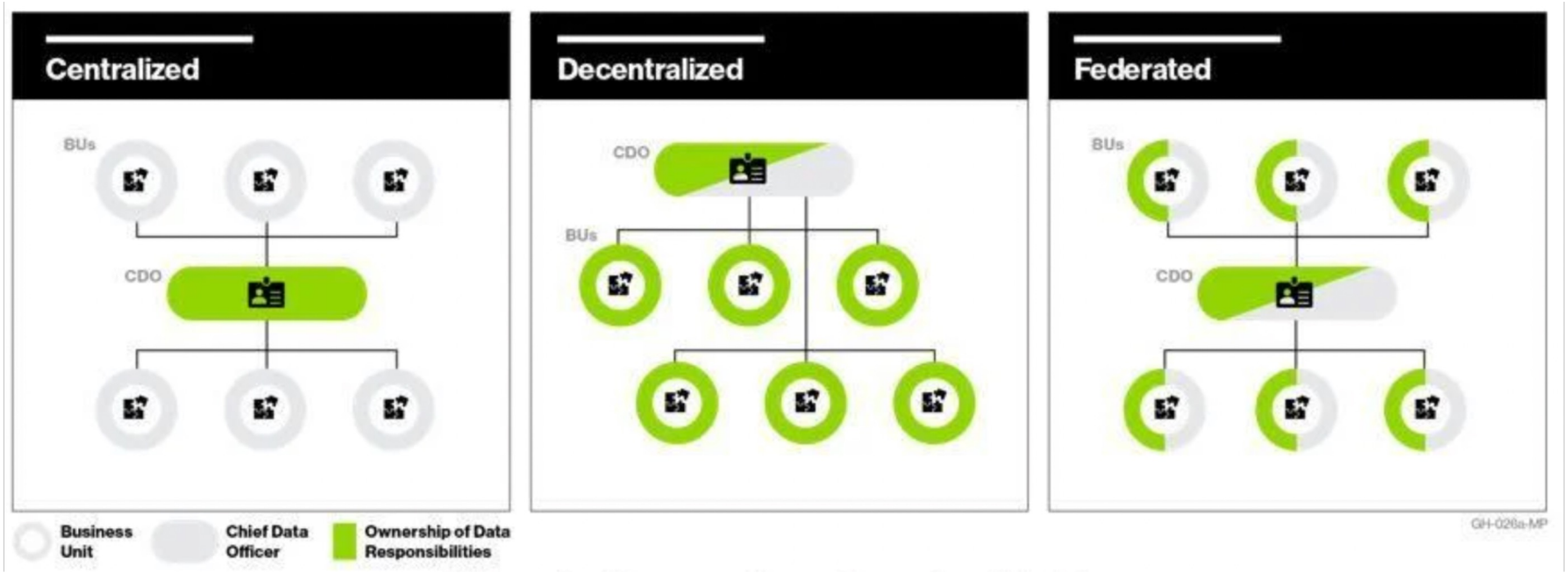
Проблема использования данных

- Проблема интерпретации данных из-за различного понимания сути показателей
  - **ПРИМЕР**: Розничная цена, событие сборка товара
- Проблема сопоставления данных
  - **ПРИМЕР**: Чеки магазинов в учетной системе и в системе лояльности
- Дублирование работ
  - **ПРИМЕР**: Детализация продаж для розницы как часть управленческой отчетности и отдельно как отчетность для сотрудников
- Проблема разделения ответственности между командами
  - **ПРИМЕР**: Учет создается разными командами без гарантии соответствия общему учету

Концепция

Выделение кроссфункциональных дата-команд, закрепленных за бизнес-доменами, сфокусированных на бизнес-ценностях, работающих по единому процессу, управляемые общими политиками и обслуживаемые единой командой дата-платформы

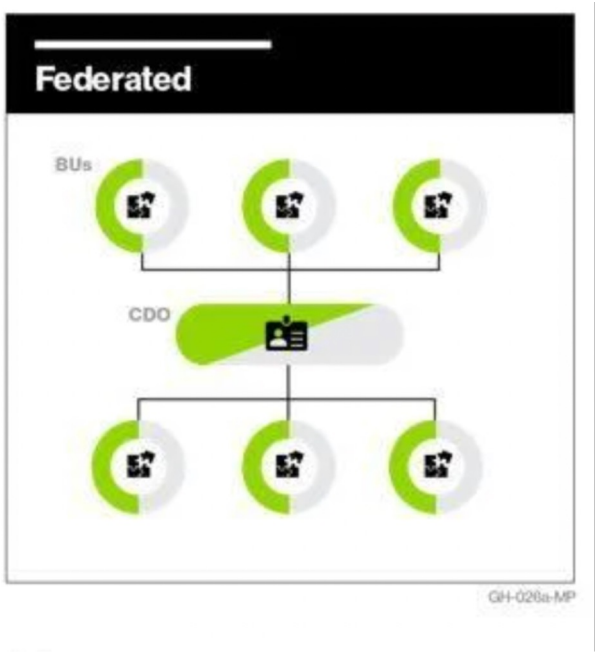
Модели структур работы с данными



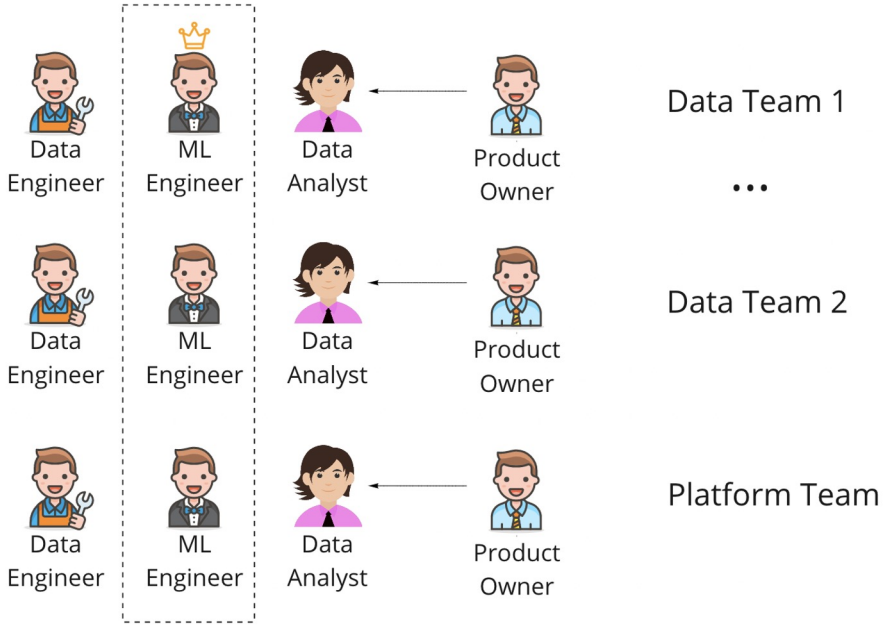
https://www.cdomagazine.tech/cdo_magazine/topics/change_management/creating-an-operating-model-to-manage-the-data-storm/article_4e9113c6-193b-11eb-a2c9-ef74a598de4e.html

Особенности федерализированной структуры

- Независимые дата-команды, закрепленные за бизнес-доменами
- Команды могут находится где угодно в структуре организации
- Команды отвечают за дата-продукты
- Кроссфункциональный состав команд (ИТ+Бизнес+Аналитики)
- Центральный Data Office обеспечивает сервисы командам
  - Data Governance
  - Self Service Data Platform
    - Общий подход к дата-архитектуре
- Общие процессы
- Общее информационное пространство
- Правила взаимодействия команд
- Правила разделения ответственности



Структура команд



Дата-команды

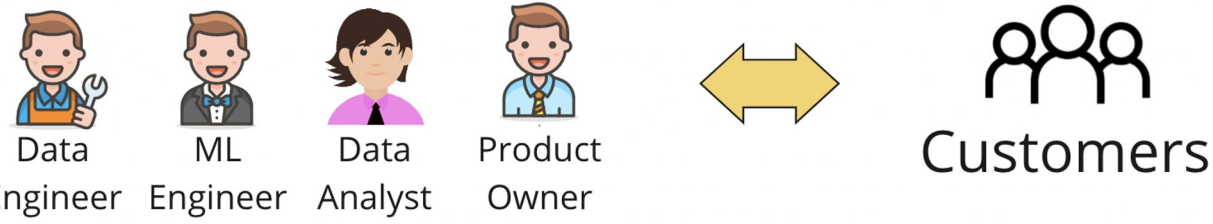
- Разрабатывают и поддерживают дата-продукты
- Отвечают за бизнес-эффект дата-продуктов
- Отвечают за инциденты тех дата-продуктов, которые поддерживают
- Inner Sourcing. Продуктовые команды тратят процент времени на развитие платформы

Платформенная команда

- Сервисная по отношению к продуктовым командам
- Обеспечивает Self-service платформу работы с данными
- Занимается общими эпиками для улучшения платформы
- Помогает командам договорится по подходам к работе с данными (безопасность, комплайенс, доступы и т.д.)

Дата Продукт Оунер

- Ориентирован на бизнес-цели
- Отвечает перед заказчиками за бизнес-результаты
- Обладает ресурсами — напрямую работает с дата-командой
- Отвечает за приоритезацию дата-эпиков
- Отвечает за оценку бизнес-эффекта дата-продуктов
- Отвечает за Жизненный Цикл продуктов (может убить дата-продукт)
- Отвечает перед заказчиками, включая другие дата-команды



Дата-домен

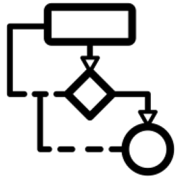
- Что такое дата-домен
  - Одна предметная область
  - Как правило, несколько связанных одной целью информационных систем
  - Внутри дата-домена общий язык (общие термины)
  - **Примеры**: складской учет, управление поставками, категорийный менеджмент
- В дата-домене могут быть несколько дата-команд
- Одна дата-команда может работать в нескольких дата-доменах
- В бизнес-домене есть один или несколько дата-доменов

Дата-продукт

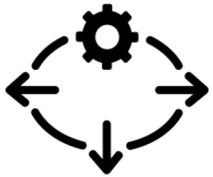
Данные - это тоже продукт
(**data as a product**)

Сырые данные

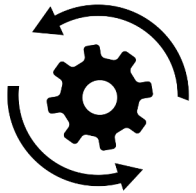
Производные данные (витрины)



Алгоритмы (производят услугу, информацию или инсайт)



Поддержка принятия решений (дашборды, отчеты)



Автоматическое принятие решение (решение принимает алгоритм/модель)

Дата-продукт - это продукт, основной целью которого является использование данных для достижения его конечной цели

Свойства дата-продуктов

- Discoverable (Обнаруживаемый)
- Addressable (Без участия команды можно найти где находится)
- Self-describing and interoperable (описание данных вместе с данными, соглашения о наименованиях, сэмплы, примеры)
- Trustworthy and secure (проверки качества данных, разграничение доступа)

Продуктовая работа для Data as a Product

- Data as a Product является дата-продуктом
- **Заказчиком дата-продукта может быть несколько других дата-команд**
- Сбор и анализ требований
  - Требования формируются вместе с заказчиками
  - Если заказчиков несколько, требования нужно обобщить
- Есть SLA (+SLO): где расположен, как получать доступ, уровень доступности, гарантии по скорости работы и пропускной способности, частота обновления, качество
- Проактивный контроль качества: ручные и/или автоматизированные проверки соответствия документации и SLO
- Реакция на инциденты (часть SLA): кто, когда, за какое время реагирует если что-то не так
- Исправление багов: гарантии (возможно, нежесткие) по времени исправления ошибок

Зачем нужна дата-платформа

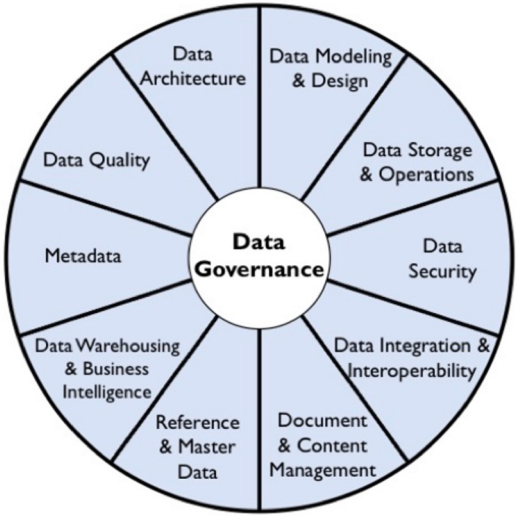
- Цели дата-платформы
  - Снижение когнитивной сложности создания дата-продуктов
  - Ускорение (Time to Market) создания дата-продуктов
- Показатель эффективности: удовлетворенность дата-команд
- У дата-платформы есть дата продакт оунер
  - Несет ответственность за платформу как продукт для дата-команд
  - Обладает ресурсами для реализации

Data Governance Council (Virtual Team)

- Виртуальная команда Data Product Owners или их делегатов
- Команда проводит регулярные встречи
- Создает Правила (Policies), определяющие принципы и регламенты совместной работы Дата-команд
- Пополняет общий бэклог работ для платформенной команды
- Правила являются обязательными для дата-команд
- Правила автоматизируются платформенной командой по мере необходимости

Data Governance

- Идентифицируем наиболее болезненные участки
- Начинаем с простых инструментов (WIKI, Excel)
- Институализируем практику использования в организации.
  - Инструмент понятен, несет ценность, большинство людей его используют, инструмент совершенствуется
- Автоматизируем или ищем готовые продукты когда понимаем ограничения и цели улучшения
- Опыт показывает, что простых инструментов может хватить на годы



https://www.dama.org/cpages/body-of-knowledge

Economy of Scale, LLM и очередная революция

- (Dis)economies of Scale — разработка ПО не масштабируется. LLM могут это изменить.
- Copilots, chatGPT, Bing…
- Просто добавь контекст

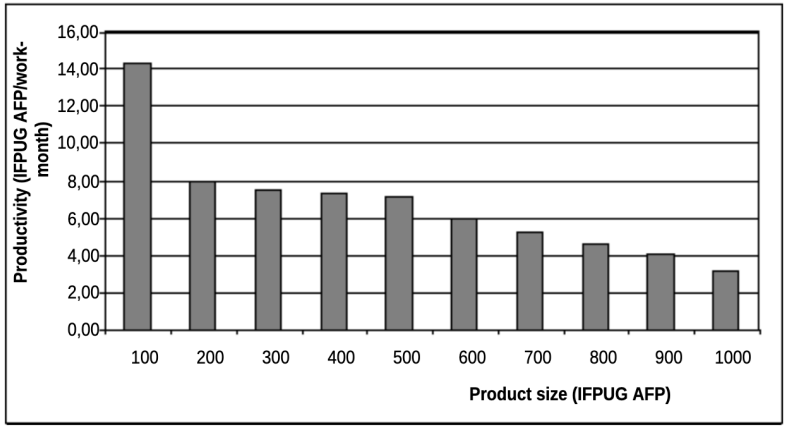
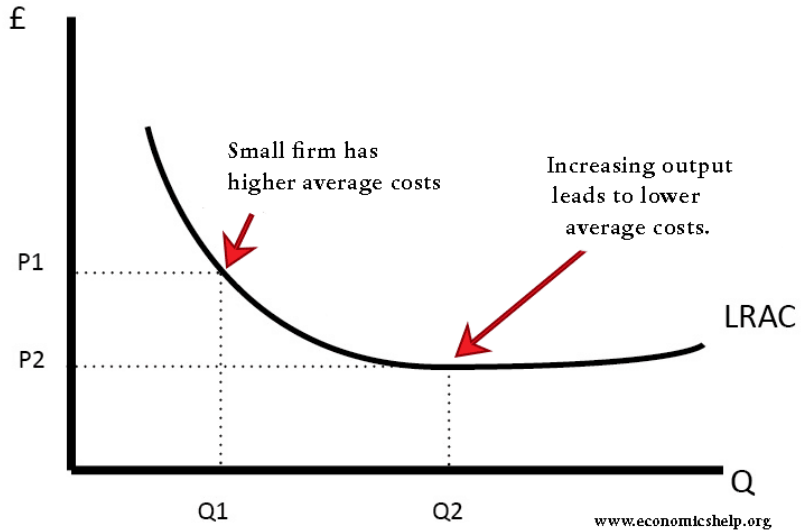


Fig. 3 BSS D&EP productivity versus product size expressed in IFPUG adjusted FP for banking software system.

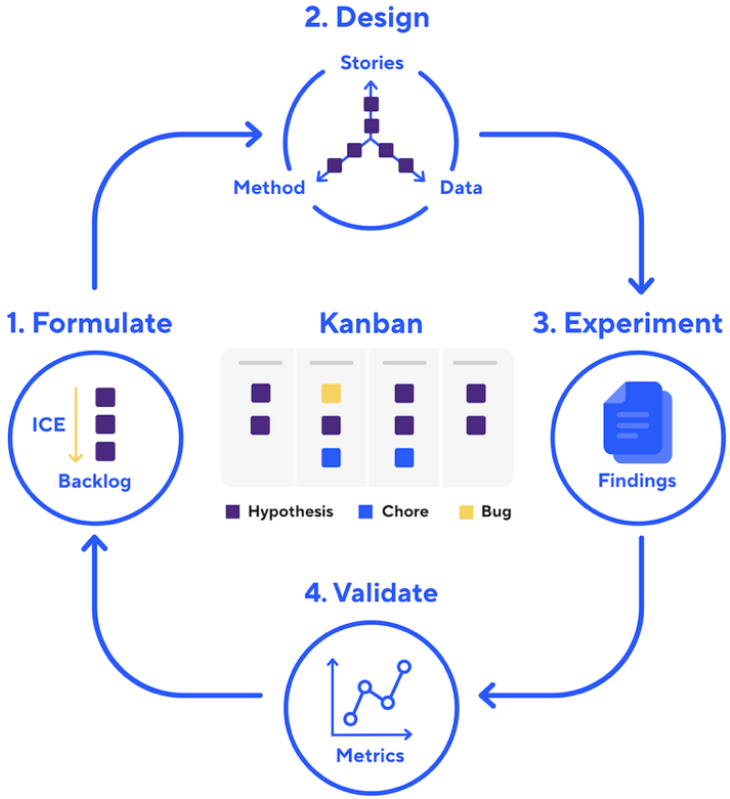


Подход LeanDS

LeanDS — открытый подход к управлению DS проектами и продуктами —

- Ориентированный на бизнес
- Основанный на гипотезах
- Командный и итеративный
- Строгий и воспроизводимый

Позволит увеличить вероятность успеха за счет прозрачности процессов, снижения Time-to-Market и фокуса на инженерном совершенстве



В компании есть чат, в котором операторы отвечают на вопросы клиентов. Пусть это робот делает?

Супер! Мы посмотрим на имеющиеся данные, посмотрим на SOTA (State-of-the-Art) и начнем строить модель!

Что может пойти не так?

ROI проекта может оказаться отрицательным

CRM команда может
отказаться интегрировать модель

У бизнеса могут быть
раздутые ожидания от
проекта

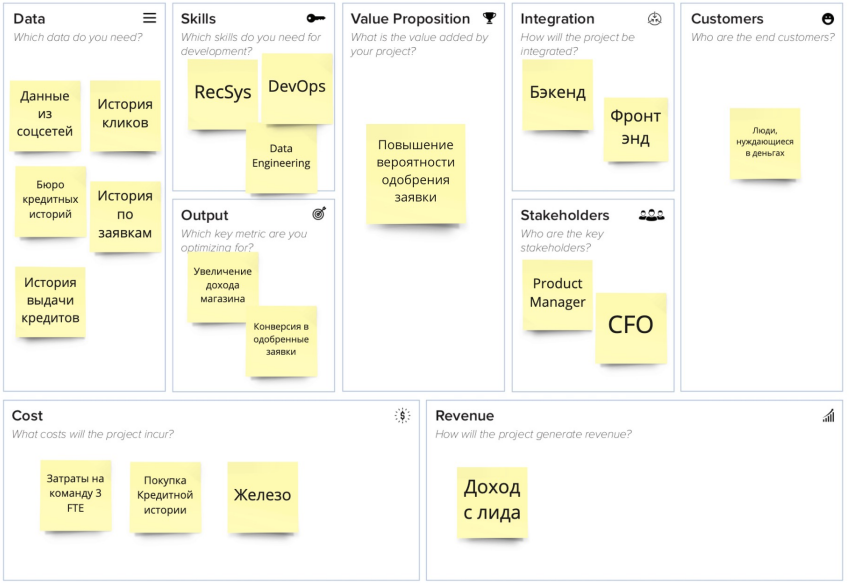
Директор колл-центра
откажется вносить
изменения в бизнес-
процесс

Top reasons

- Раздутые ожидания
- Недостаток понимания бизнеса
- Недостаток обратной связи от экспертов предметной области

AI canvas

- Одна страница
- Удобно для совместного обсуждения людей с разными компетенциями
- Фокусирует на ключевых рисках
- Быстрый способ стартовать



https://towardsdatascience.com/introducing-the-ai-project-canvas-e88e29eb7024

https://youtu.be/frtCsg8tPp4

PLANNING

Bert-Bayes
slcs-1300-cosine-distance-classifier
slcs-1299-no-glossary-intent-classifier
x1retail-cosine-glossary
btv-cosine-glossary
FastText from scratch
FastText from pre-trained embeddings
Random forest
Bert-Bayes with tuned BERT (RuBankBERT)
FastText from bank logs (dim **100**, **ns**)
FastText from bank logs (dim **300**, **ns**)
FastText from bank logs (dim **300**, **hs**)
FastText from all dialogues (dim 300, ns)
skip-thoughts логистическая регрессия
skip-thoughts полносвязная нейросеть без подбора параметров
skip-thoughts полносвязная нейросеть с подбором параметров

in Data Science project

Постановка задачи

Имплицитная постановка задачи

- Нет бизнес ценности,
- Невозможно оценить,
- Нет критериев окончания,
- Невозможно приоритизировать,
- Невозможно валидировать,
- Не понятна бизнесу, заинтересованным лицам, да и команде не всегда понятна.

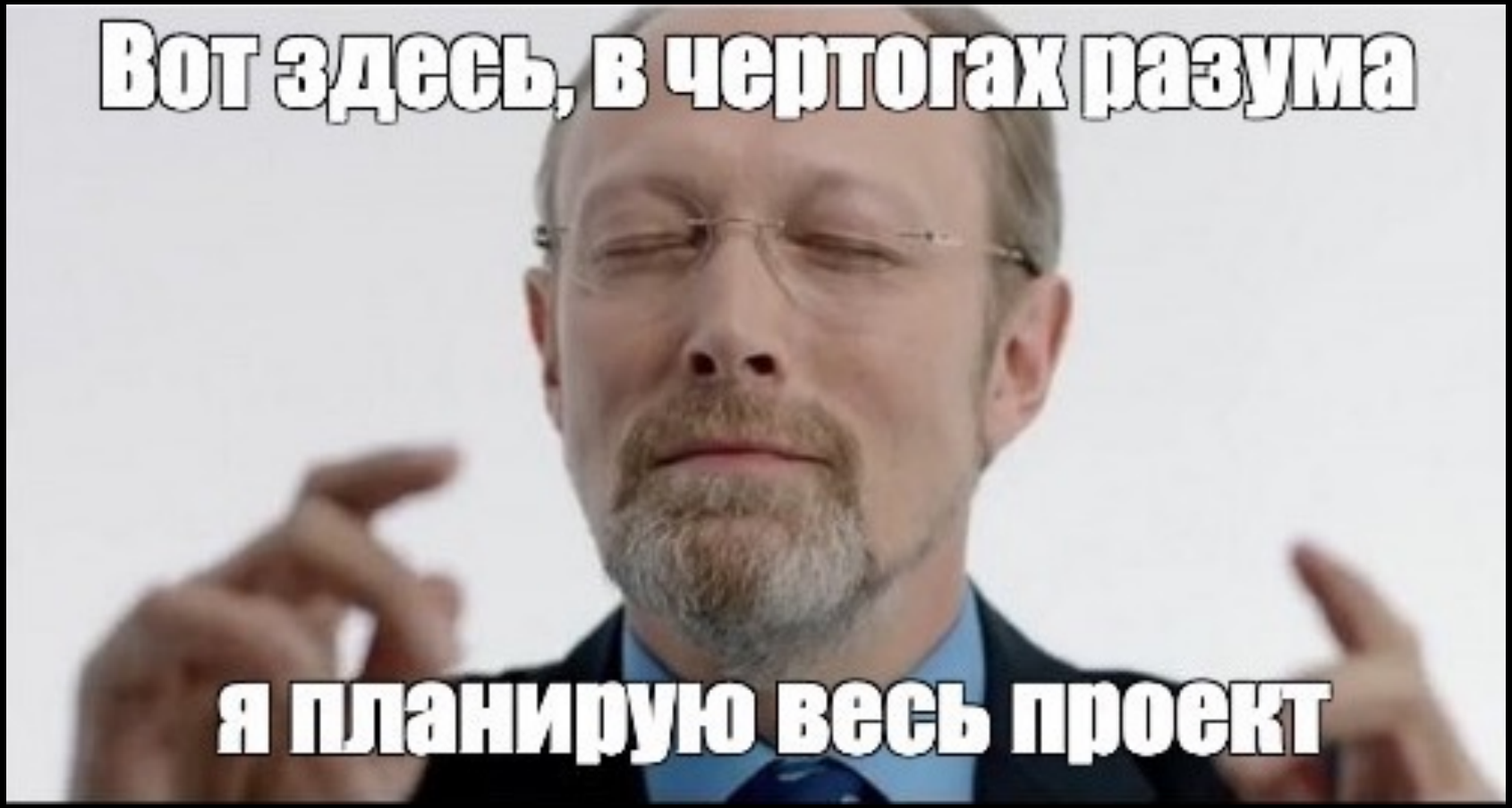
Исследовать диалоговые системы для чат.бота

Debug KNN model, которая подсказывает ответы агентам

Шаблон продуктовой гипотезы

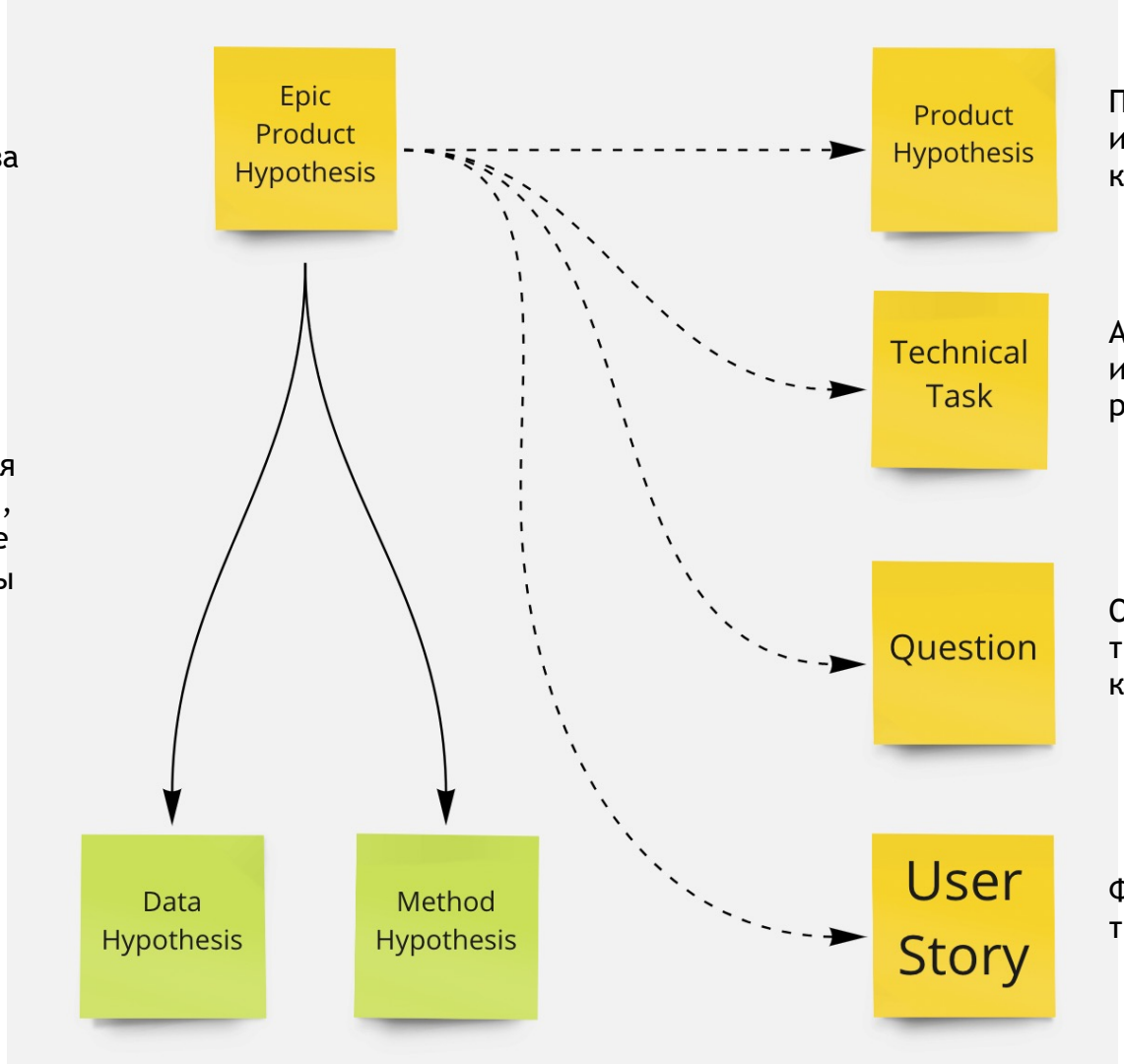
Мы полагаем, что снизим затраты на агентов, обслуживающих массовый сегмент. Для этого мы будем автоматически классифицировать клиентов. Мы окажемся правы, если увеличим количество клиентов, обслуживаемых одним агентом, которые остались довольны ответом агента на 10%.

Мы полагаем, что [решим проблему] для [аудитория]. Для этого мы [изменения]. Мы окажемся правы, если получим [метрики]


Вот здесь, в чертогах разума
я планирую весь проект

Эпик — большая продуктовая гипотеза

Эпик декомпозируется на небольшие работы, в том числе на другие продуктовые гипотезы



Продуктовая гипотеза имеет ценность для клиента или бизнеса

Архитектурная или инфраструктурная работа

Открытый вопрос, требующий времени команды

Функциональные требования

Data/Method гипотеза

Мы полагаем, что на кластеризованных вопросах пользователя и ответах агента можем построить линейную модель классификации для выбора подсказки. Это позволит получить Top 5 Accuracy не менее 50%.

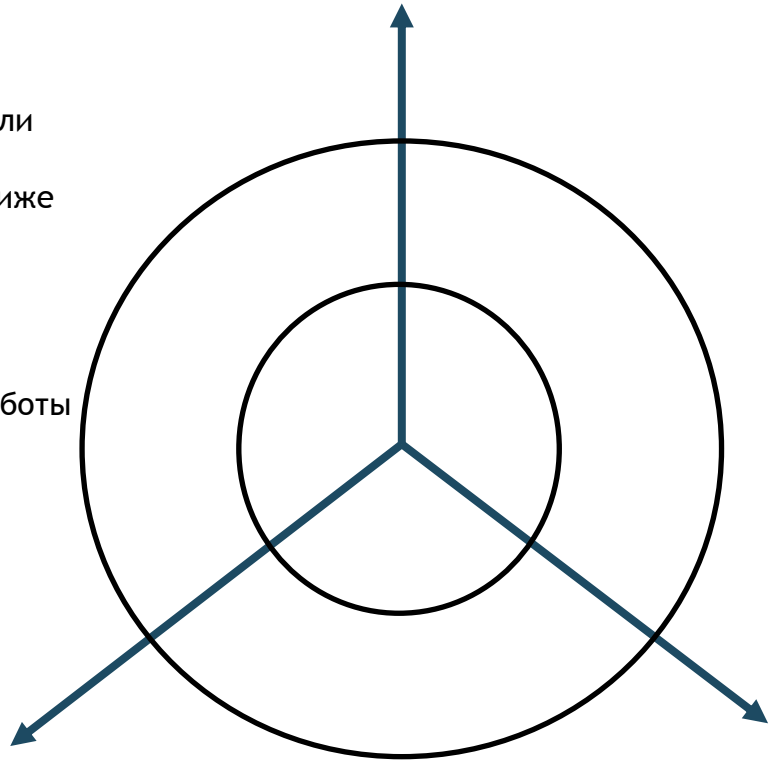
Мы полагаем, что [используя данные] и [метод машинного обучения] мы получим [метрику] не меньше [значение]

Брейнсторминг гипотез

User Stories
Функциональные требования

Questions
Если есть сомнения или вопросы — добавляем стикер с вопросом ближе к его источнику

Technical Tasks
Архитектурные и инфраструктурные работы



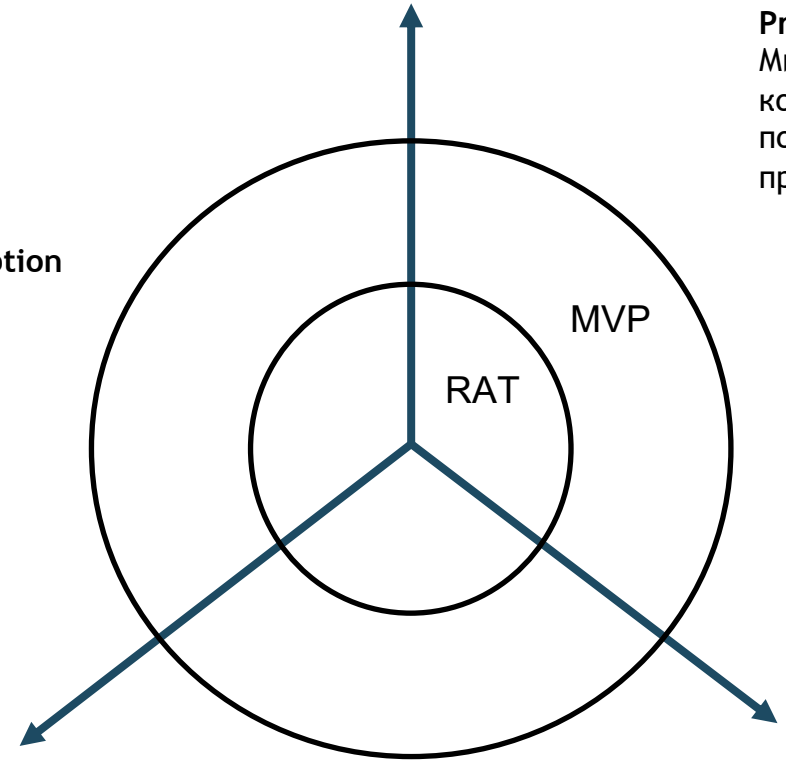
Method Hypotheses
Гипотезы про использование данных

Data Hypotheses
Гипотезы про использование данных

Брейнсторминг гипотез

MVP (Minimum Viable Product)
Минимальный продукт, который можно показать пользователю чтобы проверить гипотезу

RAT (Riskiest Assumption Test)
Самые рискованные предположения



Data

1.Достаточно ли данных для обучения модели?
2.Есть ли в данных для обучения предсказательная сила для решаемой задачи?
3.Качество данных достаточно для использования в модели?
4.Есть ли разметка?
5.Можно ли купить дополнительные данные?
6.Можно ли использовать публично доступные данные?
7.Заражены ли данные предрассудками (biased)?
8.Как модель будет получать данные в проде?
9.Совпадают ли данные в трейне и проде?
10.Как контролировать качество данных в проде?
11.Как сохранить качество данных после выхода в прод?
12.Какие юридические риски связаны с использованием данных?
13.Есть ли какие-то требования к данным, связанные с законодательным регулированием?

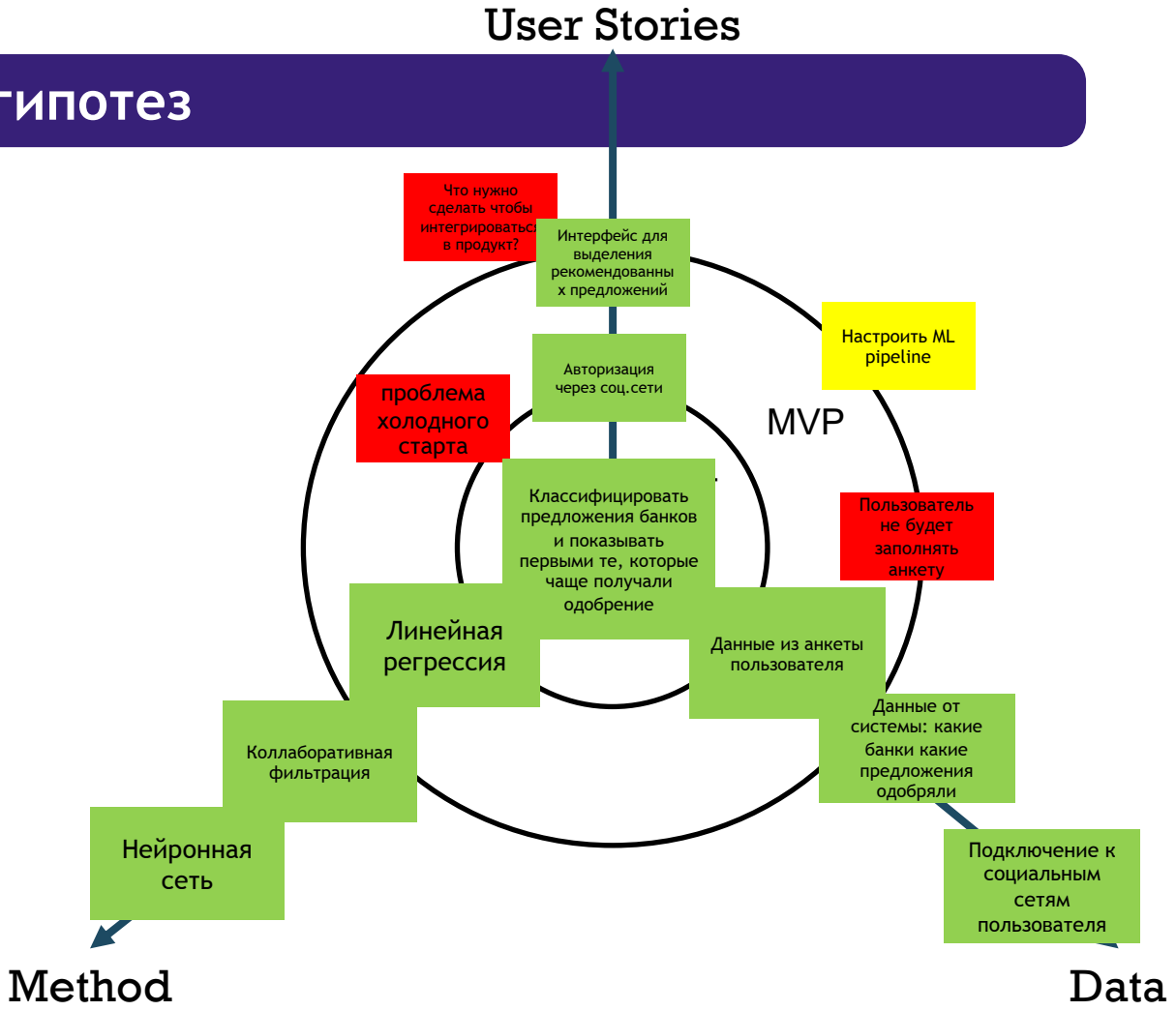
Story

1.Как будет работать сервис, когда когда модель выйдет в прод?
2.Какие технические ограничения существуют для сервиса?
3.Какие системы и команды зависят от сервиса?
4.Как проводить валидацию модели в проде (А/Б тест, ...)? Как валидация должна быть устроена?
5.Что будет, если сервис откажет?
6.На сколько быстро меняются данные, Как частно нужно повторно обучать модель?
7.Как будет производится мониторинг работы модели и сервиса в проде?

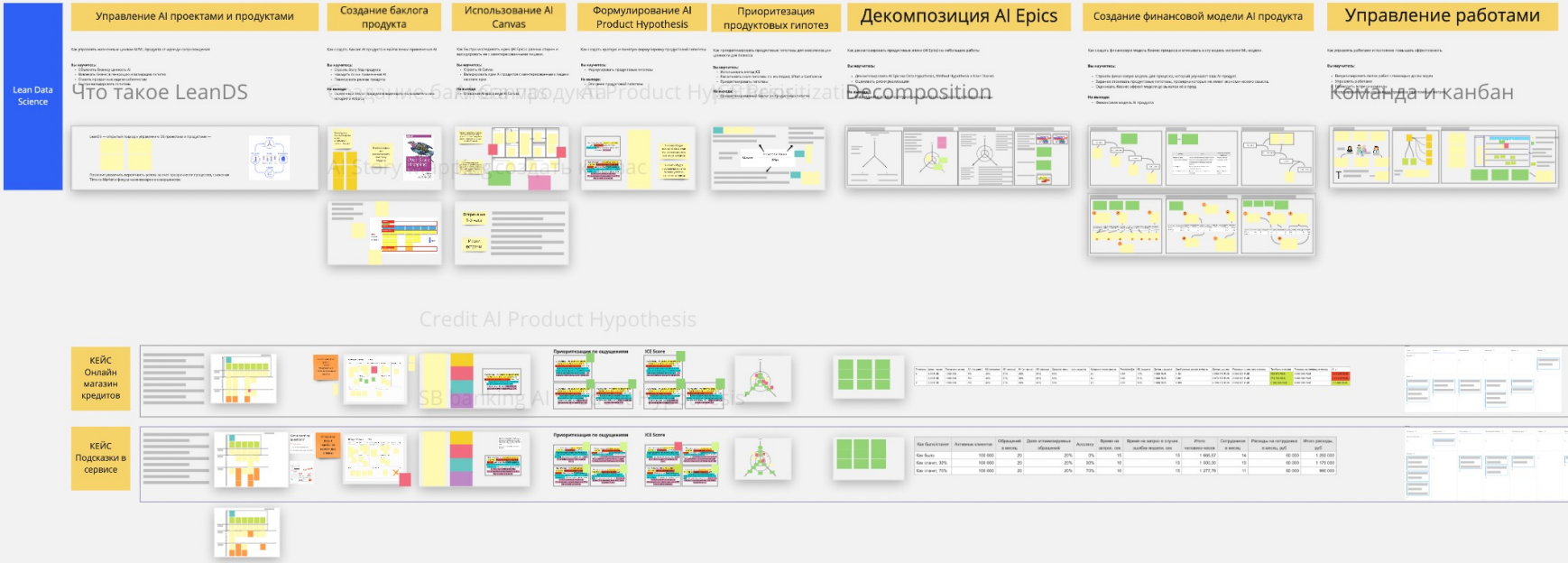
Method

1.Есть ли SOTA (State of the Art) для задачи? Какие у нее значения мер качества?
2.Как эту работу делают эксперты (Subject Matter Expert) сейчас?
3.Требуется ли интерпретируемость модели?
4.Требуется ли внешняя независимая валидация модели?

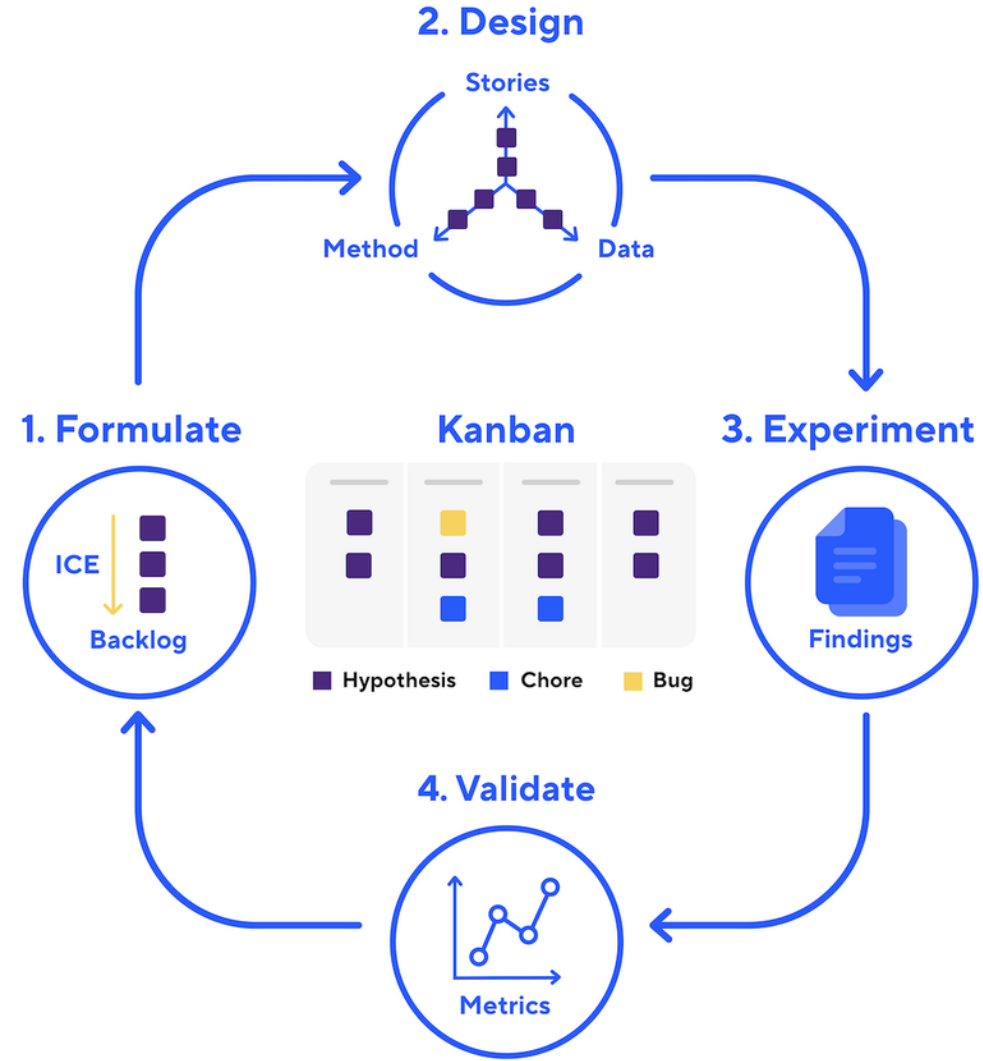
Брейнсторминг гипотез



LeanDS Cheat Sheet



https://cntr.click/vv6CaQV



Lean Data Science

- @leands
- leands_ru
- Http://leands.ai
- http://leands.university/

Асхат Уразбаев @askhatu