Хранилище есть, а дальше что? Документация и другие способы улучшить DX ваших коллег

Игорь Мосягин

«Я инженер|аналитик|менеджер, меня никто не понимает, памагити»

ПРО ЧТО

представьте такую ситуацию

- есть небольшая команда аналитиков, они себе что-то ковыряют
- есть большая команда инженеров, в том числе часть из них помогает аналитикам катить в прод их идеи
- time to market очень большой
- аналитики недовольны тем, как работают инженеры
- инженеры недовольны тем, как пишут код аналитики
- менеджеры обеспокоены что нельзя получить внятный результат (хотя данных аж 500 терабайт)

Знакомо? Давайте поговорим об этом

о чём будет доклад

- чего хотят инженеры
- чего хотят аналитики
- чего хотят менеджеры
- как так вышло, что кто-то не понимает друг друга
- что с этим можно сделать сейчас
- что с этим можно сделать вообще

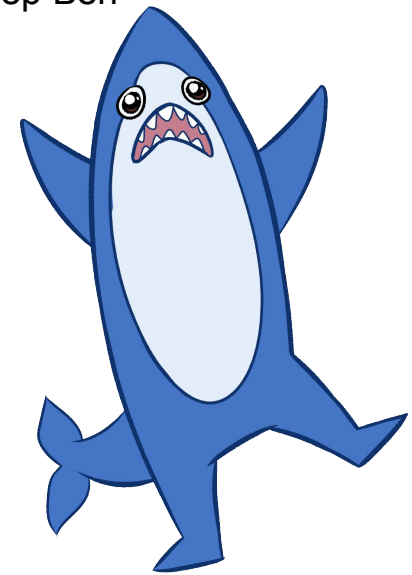
а я кто

- писал бэкенд
- делал сайнс
- делал дата-сайнс
- сейчас инженерю данные в международной финтех-компании Klarna
- хожу на дс-хакатоны поэтому не забыл как там у аналитиков,
  пока там ничего не поменялось и там бардак
- общаюсь с другими инженерами и сам инженер,
  поэтому знаю какой бардак и тут
- менеджерить не люблю, но немножко иногда приходится

кто мне будет помогать с докладом

- да кто такой этот аналитик, что это за зверь инженер?

Инженер Бен



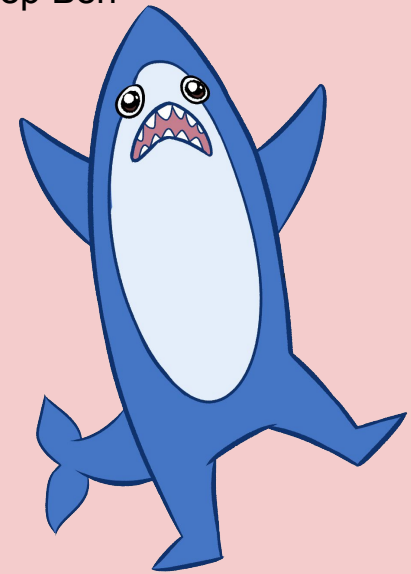
Менеджер Рикардо

Аналитик Агафья

кто мне будет помогать с докладом

- да кто такой этот аналитик, что это за зверь инженер?

Инженер Бен



не работал по специальности, сразу пошёл писать сайты

дома есть 6 ардуино и 3 raspberry pi (пылятся)

имеет на всё своё мнение:
какой облачный провайдер лучше
какая механическая клавиатура правильная
как готовить шакшуку

настолько давно пишет код,
что смеётся с этой шутки ➡

кто мне будет помогать с докладом

- да кто такой этот аналитик, что это за зверь инженер?

Аналитик Агафья

пока писала кандидатскую по биологии, понравилось ковырять данные больше, чем писать статьи

знает “язык программирования” R, учит python и обладает интуитивным чутьём к приятным цветовым гаммам на графиках

понюхав пороха настоящей науки, не любит термин “дата-сайнтист” и зовёт себя “аналитик”

кто мне будет помогать с докладом

- менеджер самый понятный из всех этих зверей

Менеджер Рикардо

заслуженно считает себя очень умным
не писал код после второй работы, но иногда пописывает в стол что-то

говорит мало, но по делу
обожает гугл-таблички

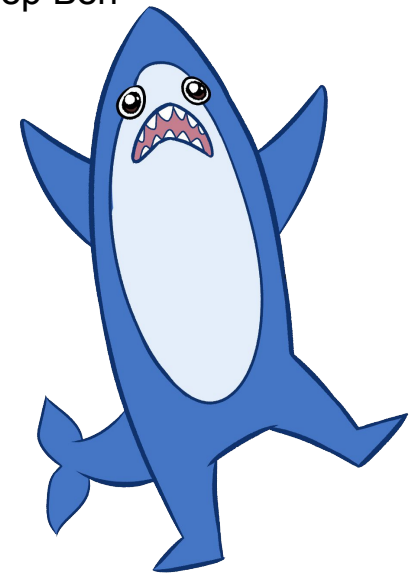
менеджер Агафьи, а Бен его боится

откуда-то понимает что такое Data Mesh, Data Lake, Data Warehouse и зачем нужен Data Governance

кто мне будет помогать с докладом

- Наши герои в ООО “Плавники и Лапки” строят e-commerce-маркетплейсы

Инженер Бен



Менеджер Рикардо

Аналитик Агафья

Менеджеры

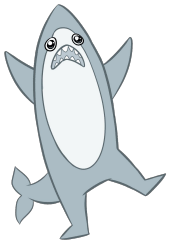
Они же accountability managers, project managers, product managers, etc.

Что делают и чего хотят менеджеры

- по сути, хотят прогнозируемости
- договариваются со всеми подряд
- лепят гугл-таблички и может тасочки в джире
- два основных инструмента это браузер и мессенджер
- знают всякие аббревиатуры: MVP, RACI, RAT
- не подают виду, но не любят когда им предлагают сделать PR
- думают о том о чём не думают обычно IC

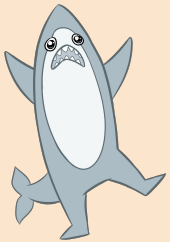
ИНЖЕНЕРЫ

могут быть просто backend SWE или же прям DE, R&D devs и подобное



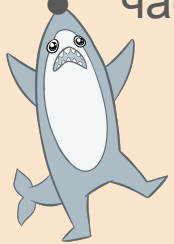
Что делают инженеры

- пишут код, понимают, как его тестировать
- используют / любят практики совместного написания кода
  - код-ревью
  - тесты
  - “пиар нельзя вмержить, потому что джоба с тестами падает, пофиксь ошибки линтера”



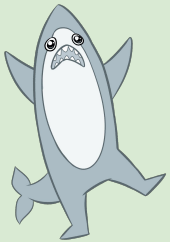
Что делают инженеры

- пишут код, понимают, как его тестировать
- используют / любят практики совместного написания кода
  - код-ревью
  - тесты
  - “пиар нельзя вмержить, потому что джоба с тестами падает, пофиксь ошибки линтера”
- имеют разные среды для тестирования local/dev/staging/prod
- часть компетенции инженера заложена в понимании инфраструктуры
- “если нет документации, то код — это документация”
- часто ими погоняет отдельный человек, который “общается с бизнесом”



Инженеры в дикой природе

- писать код в свободное время — это нормальная идея, даже весело
- могут знать другие языки или подходы, отличные от того, что они делают
- привыкли настраивать себе рабочую среду так, чтобы она помогала
  - гитхуки, линтеры, плагины к IDE
  - если есть какой-то частый раздражитель, то “наверное, есть плагин, это упрощающий”



Аналитики

Они же дата-сайнтисты, analytics engineers, data analysts, etc.

Что делают аналитики

- помогают бизнесу делать data-driven решения
- часто подгоняются бизнесом: “Давай в тест побырику”
- читают новости своей области, конференции смотрят,
  следят за статьями, новыми фишками в библиотеках и так далее
- часто не видят смысла делать то, что инженеры считают common sense
  (привет, техдолг)
- Untitled1.ipynb, Untitled3.ipynb, Untitiled_JIRA7534.ipynb…
  и они помнят где что лежит, во всяком случае первые пару недель

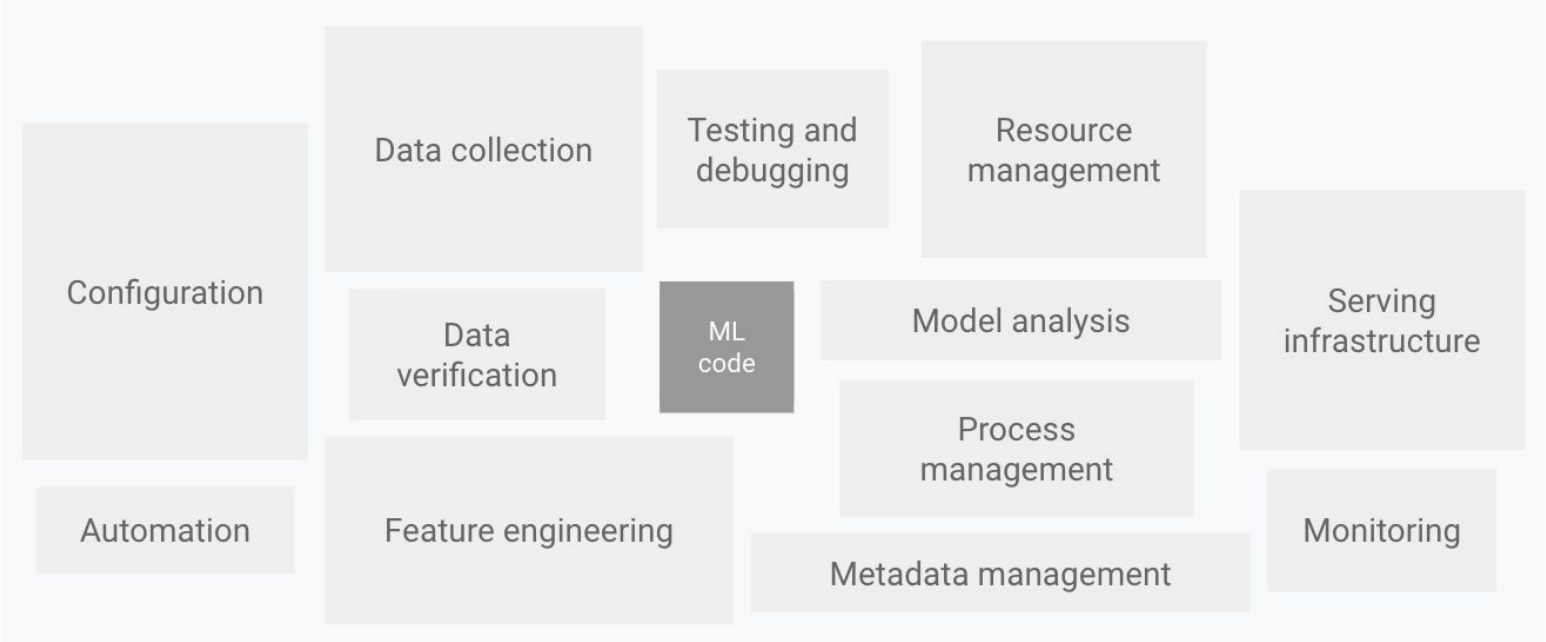
Как живут аналитики

- анонимизированные данные делать можно но сложно,
  поэтому тестируем на проде
- слышали про “докер-контейнер”, пугаются если попросить написать свой
- главное — получить хорошую метрику, неважно, какой там код
- ощущение “я закончил” заканчивается на числах в целевом датасете
  (или проанализированным а/б-тестом)

ПОЧЕМУ

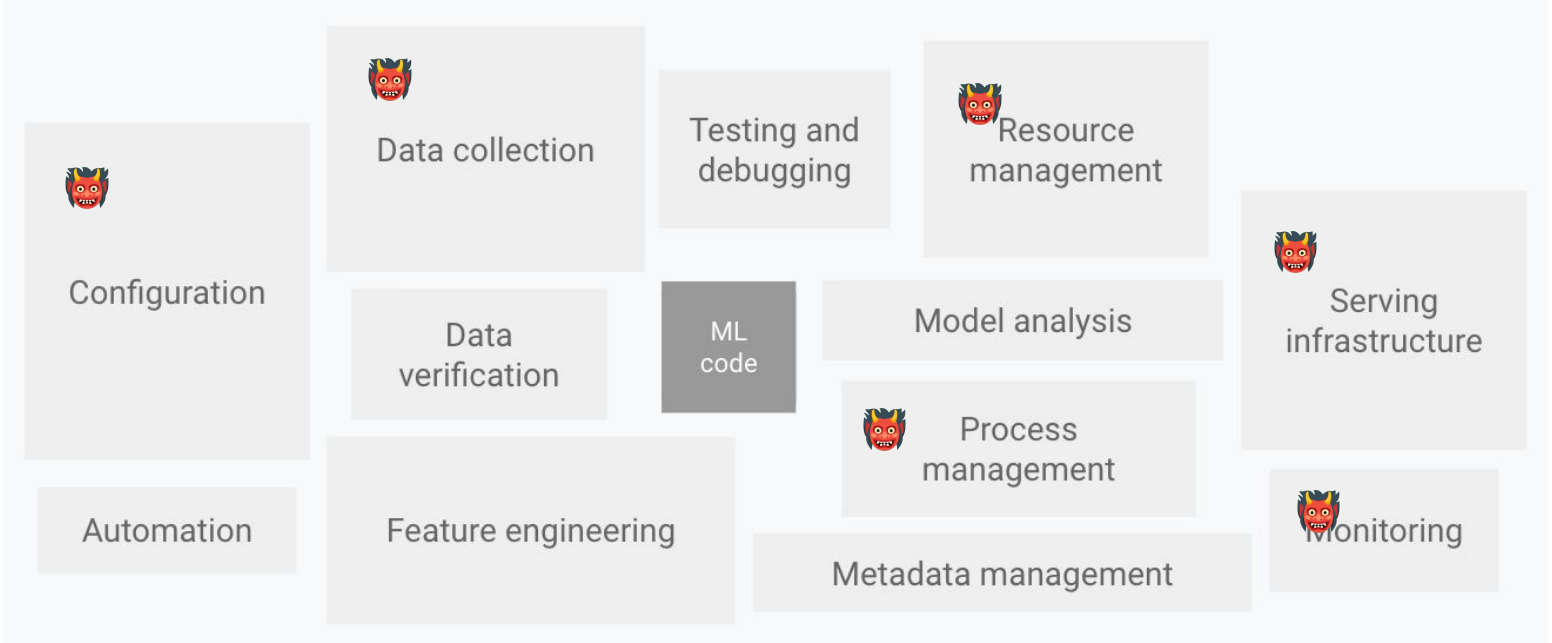
что почему? почему аналитики так отличаются от инженеров
и менеджеры могут быть обеспокоены этим

Техдолг в дата-продуктах на примере ML-систем



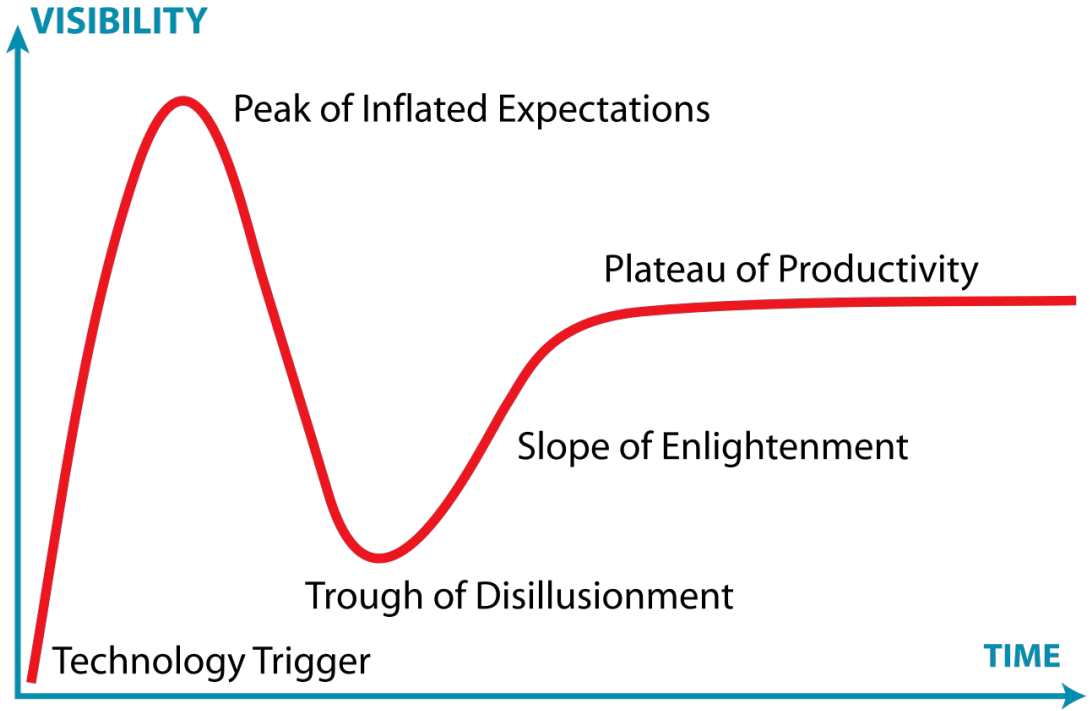
картинка из статьи “Hidden Technical Debt in Machine Learning Systems”
с тех пор инструменты стали запутаннее, данных больше, дашборды жирнее…

Этой картинке 10 лет (она всё ещё актуальна)



нужны уже специальные биг-дата-тулы просто понять, что происходит в твоей системе! Давайте посмотрим внимательнее на примере

Ну ту картинку мы уже видели а что там Gartner?



Выводы из клёвого свежего [кейнота]

- Данных становится больше, а инсайтов меньше/сложнее
- Нужен не только процесс сбора и аналитики но и governance
- Governance is the process of deciding on how to get things done done

Тут много low-hanging fruits, их можно посрывать, попробуйте их поискать

Какой-то ещё загон про книгу “Цель” который я в последний момент недоформулировал

ЗАРИСОВКА

из жизни компании “Плавники и Лапки”

менеджер просит аналитика задеплоить модель

мы в этом квартале хотели поэкспериментировать с полкой “с этим товаром покупают”, фронт говорит что у них готово, бэк тоже может заменить заглушку быстро

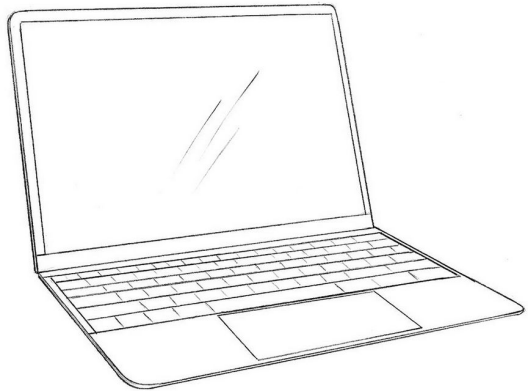
менеджер просит аналитика задеплоить модель

ещё мы переезжаем в облако всё-таки, я говорил с Беном там ничего сложного, сделаешь? Хочу этим похвастаться на слёте департаментов

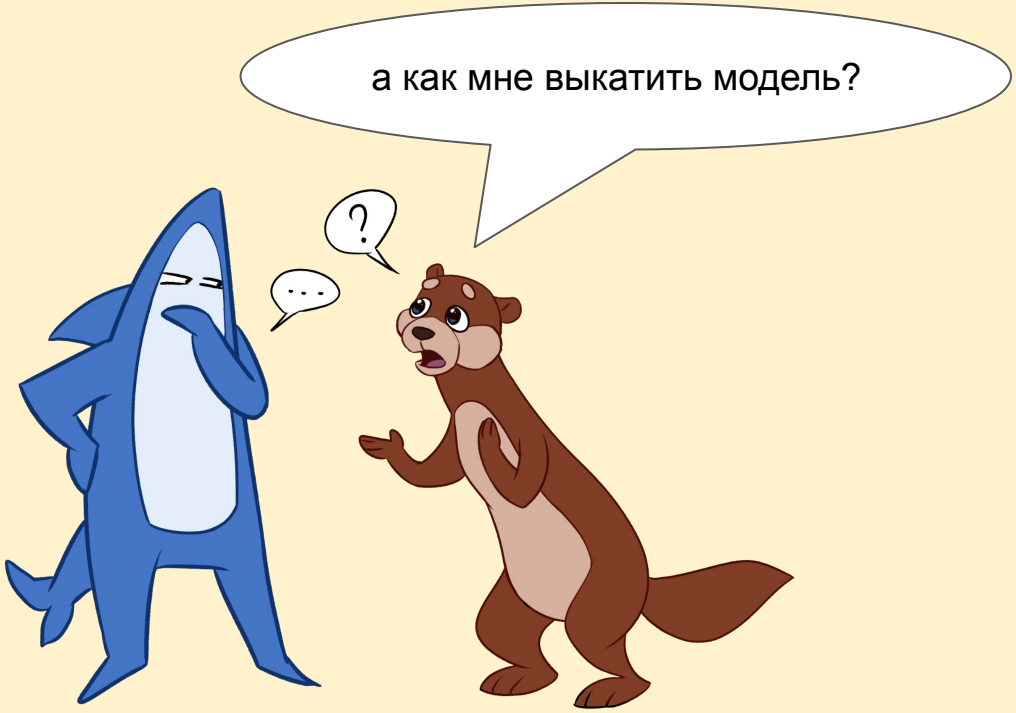
менеджер просит аналитика задеплоить модель

аналитик погружается в это всё

- модель обучается локально
- есть облачные решения и для обучения,
  но там как-то муторно и непривычно
- **как аналитику задеплоить модель в прод?**
- поищу доки на портале компании
  (хаха)

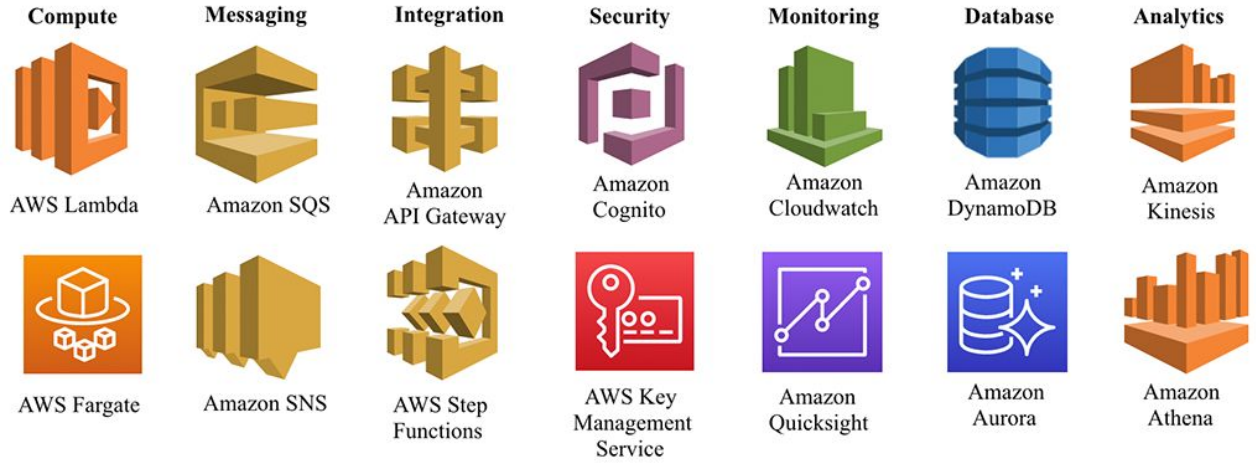


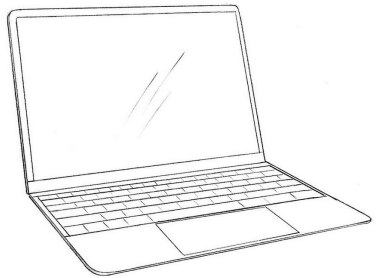
аналитик приходит за помощью к инженеру



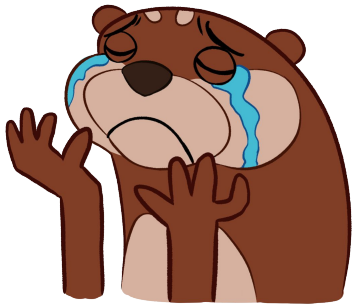
аналитик приходит за помощью к инженеру

Аналитик получает доступ и ищет, как обучить модель

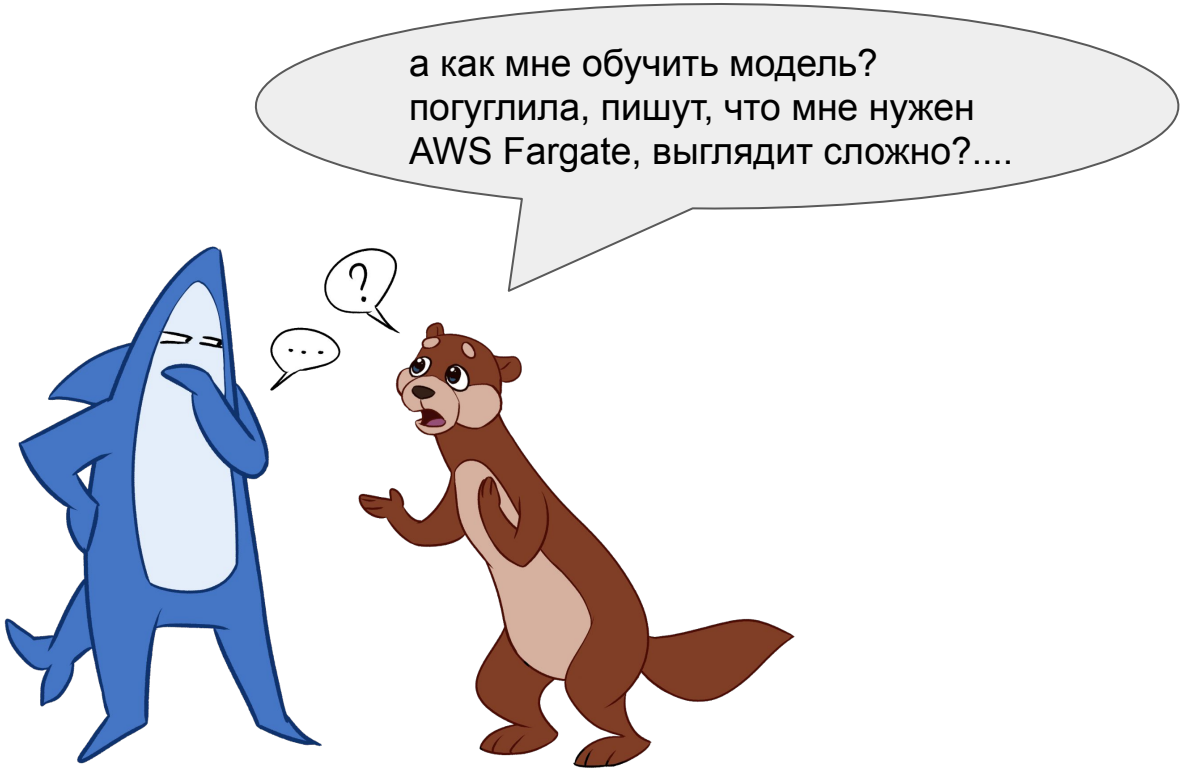




Иконки красивые,
но ничего не понятно

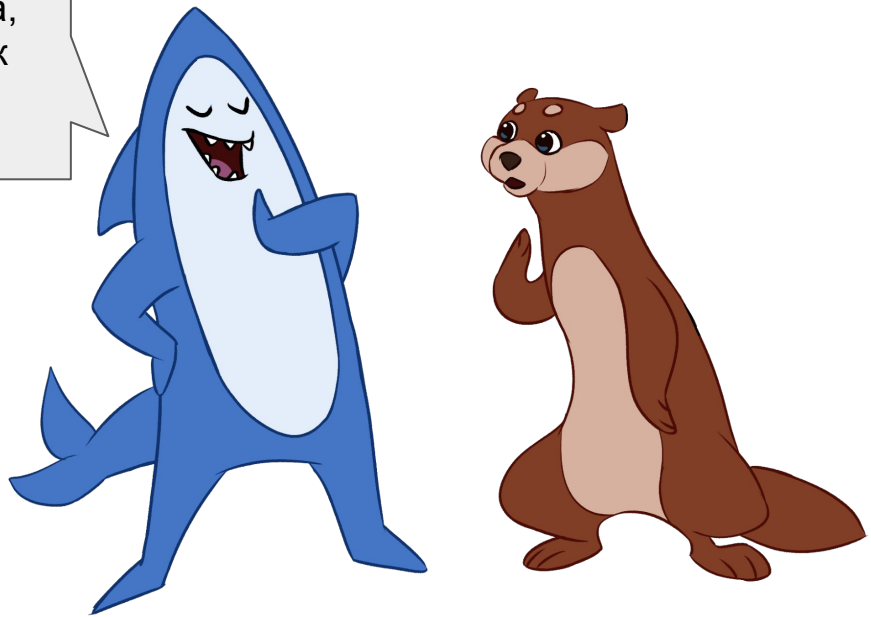


аналитик приходит за помощью к инженеру



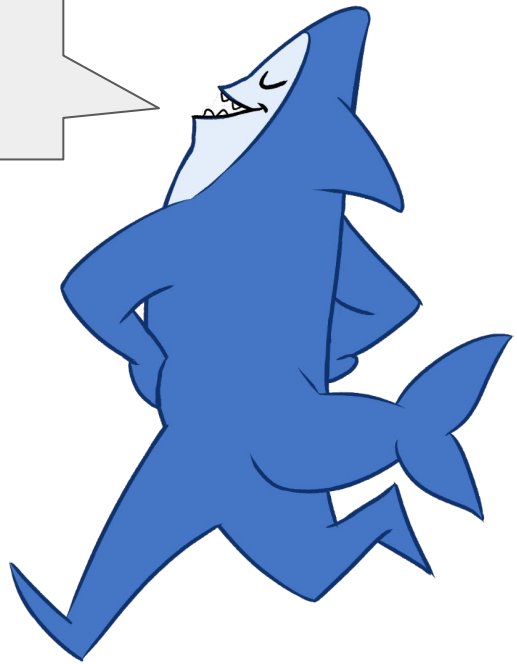
аналитик приходит за помощью к инженеру

У тебя же просто тетрадка? Для тетрадок тебе нужен Sagemaker, потом тебе нужна API Gateway и AWS Lambda, чтобы привязать эндпоинт к твоему этому, как его, инференсу



аналитик приходит за помощью к инженеру

Ну, ты пиши, если что, но там всё просто



aws | Services | Search for services, features, blogs, docs, and more [Alt+S] | Stockholm | aws-igor-ms

Console Home Info

Actions

Recently visited Info

No recently visited services

Explore one of these commonly visited AWS services.

IAM EC2 S3 RDS Lambda

Welcome to AWS

Getting started with AWS
Learn the fundamentals and find valuable information to get the most out of AWS.

Training and certification
Learn from AWS experts and advance your skills and knowledge.

What's new with AWS?
Discover new AWS services, features, and Regions.

AWS Health Info

Open issues
0 Past 7 days

Scheduled changes
0 Upcoming and past 7 days

Other notifications
0 Past 7 days

Go to AWS Health

Cost and usage Info

Current month costs
$0.00

Forecasted month end costs
$0.00

Last month costs
$0.00

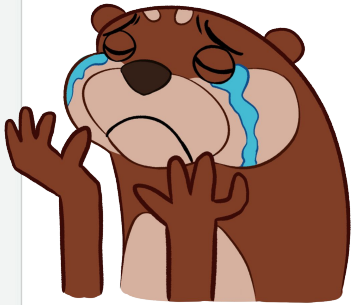
Costs shown are unblended. Learn more

Top costs for current month

No top cost breakdowns. Cost breakdowns show when you use services.

Go to AWS Cost Management

Бен сказал что просто… Но… Куда тут жать-то?



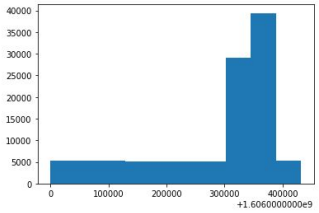
File Edit View Run Kernel Tabs Settings Help

Solver.ipynb

Code

Python 3 (ipykernel)

```
[1]: import json

[2]: skus2cat = json.load(open("_tech_2100_skus.json"))["skus"]
     users = json.load(open("_users_200.json"))["users"]

[3]: events = [json.loads(event) for event in open("_tech_events.jsonl").readlines()]

[4]: import operator

[5]: timestamps = list(map(operator.itemgetter("timestamp"), events))

[6]: from matplotlib import pyplot as plt

[7]: plt.hist(timestamps)

[7]: (array([ 5248.,  5219.,  5231.,  5182.,  5161.,  5105.,  5105., 29017.,
             39417.,  5315.]),
      array([1.60599958e+09, 1.60604280e+09, 1.60608602e+09, 1.60612924e+09,
             1.60617246e+09, 1.60621568e+09, 1.60625889e+09, 1.60630211e+09,
             1.60634533e+09, 1.60638855e+09, 1.60643177e+09]),
      <BarContainer object of 10 artists>)
```

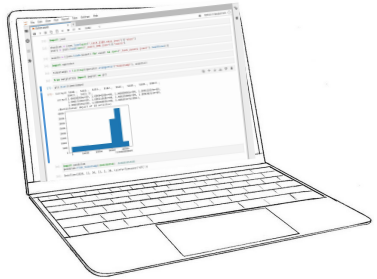


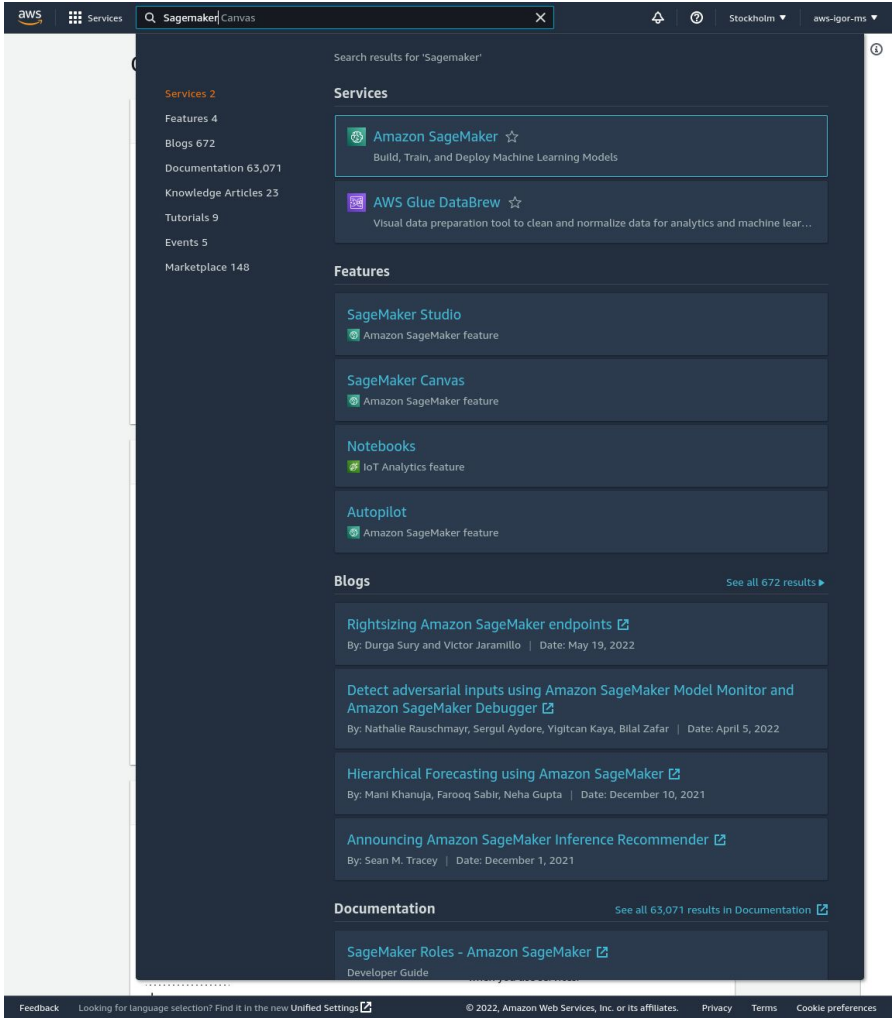

```
[8]: import pendulum
     pendulum.from_timestamp(1606388550)  #1606345330)

[8]: DateTime(2020, 11, 26, 11, 2, 30, tzinfo=Timezone('UTC'))

[9]: sku_cats = list(skus2cat.values())

[10]: import contextlib

      import psycopg2
      import pandas as pd


      with contextlib.closing(
          psycopg2.connect(
              database="sku_info",
              user="lab05",
              password="zua0ieMahk9Jei",
              host="data.ijklmn.xyz",
          )
      ) as conn:
          categories = pd.read_sql(
              """
                  select id, cat, parent_id
                  from category_tree
              """,
              con=conn,
              index_col='id',
          )
```

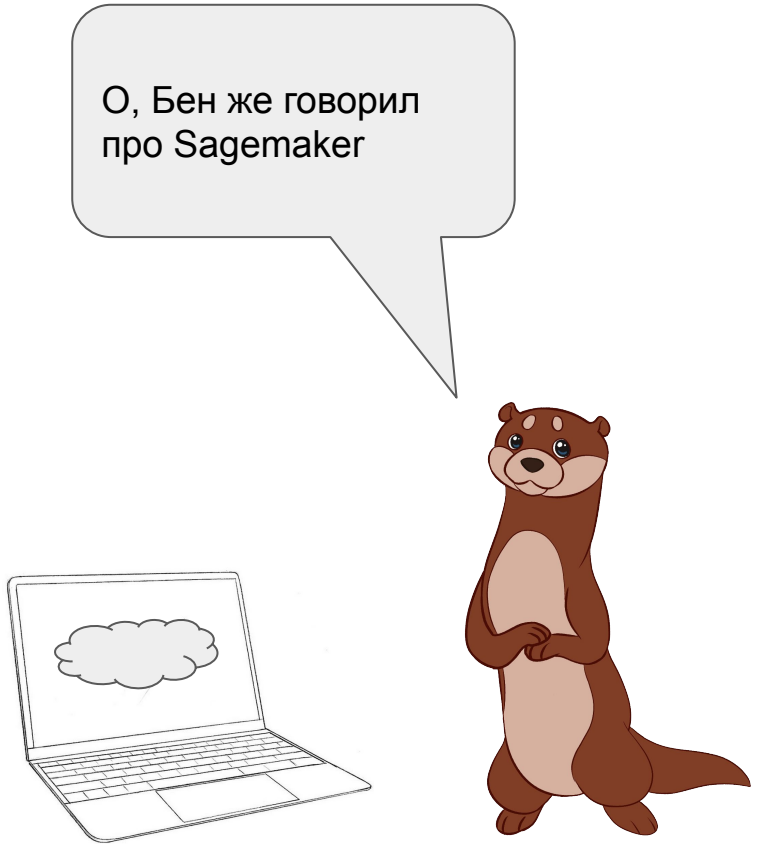
Simple 0 1 Python 3 (ipykernel) | Idle Mode: Command Ln 1, Col 1 Solver.ipynb

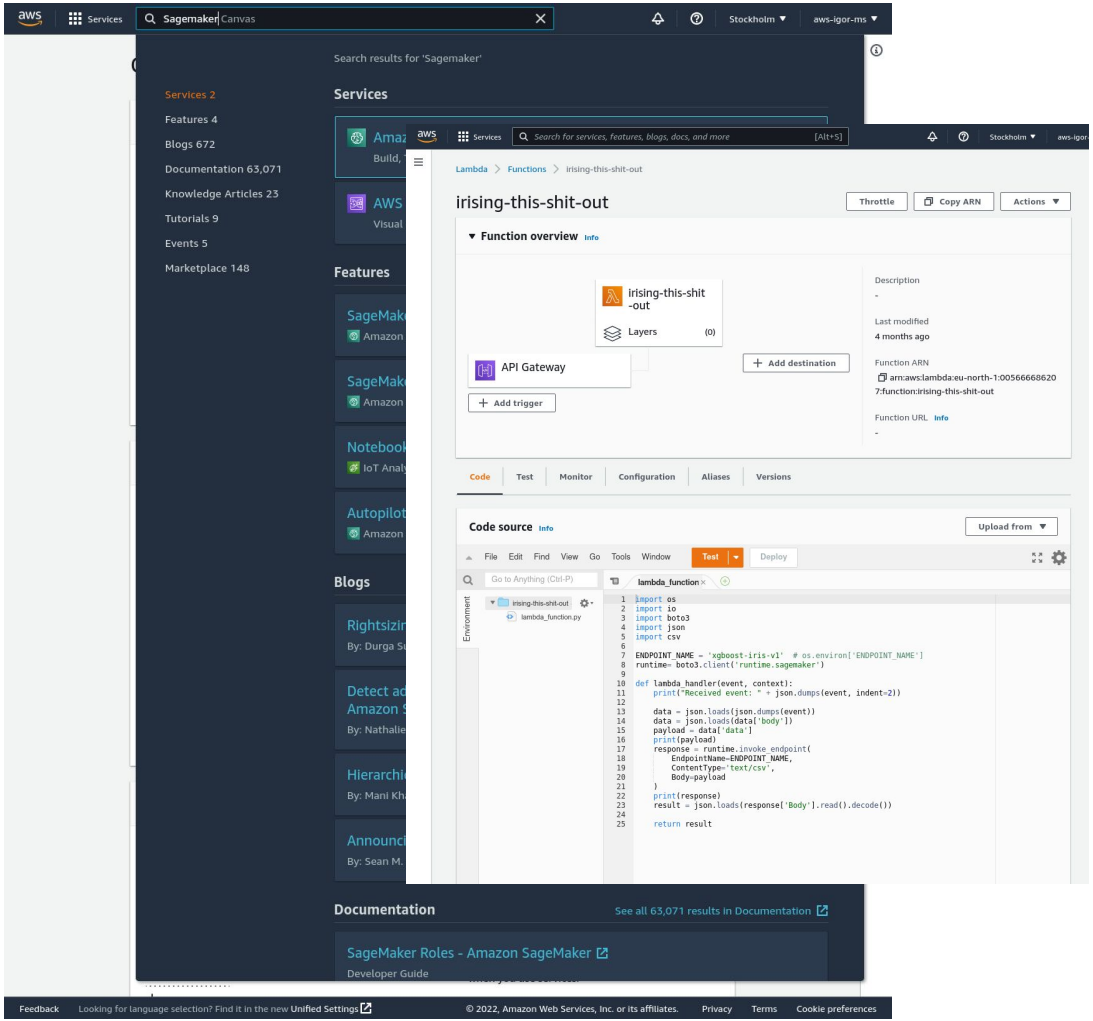
Как выглядит привычная среда для Агафьи

Даже кнопка Play есть!

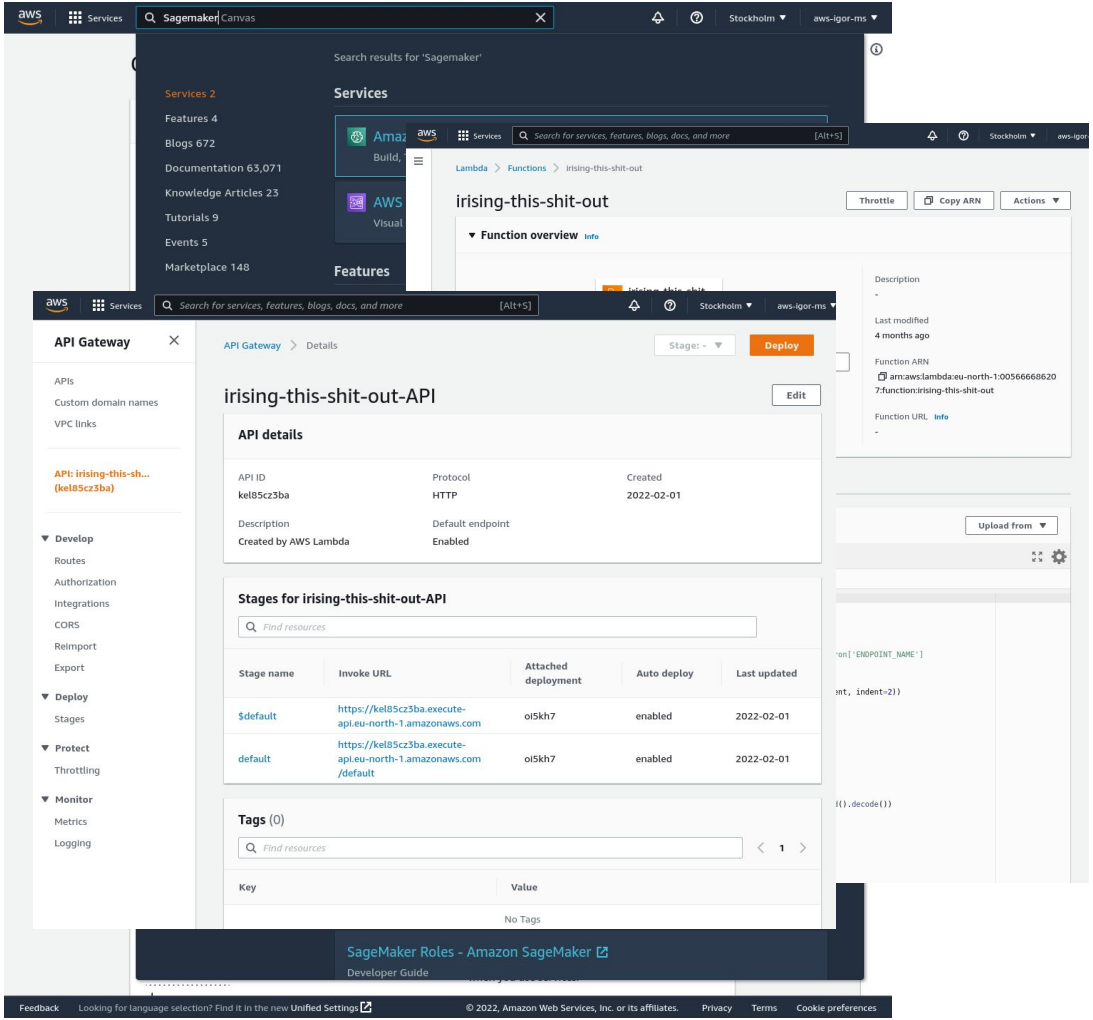



aws
Services
Sagemaker Canvas
Stockholm
aws-igor-ms
Search results for 'Sagemaker'
Services 2
Features 4
Blogs 672
Documentation 63,071
Knowledge Articles 23
Tutorials 9
Events 5
Marketplace 148
Services
Amazon SageMaker
Build, Train, and Deploy Machine Learning Models
AWS Glue DataBrew
Visual data preparation tool to clean and normalize data for analytics and machine lear...
Features
SageMaker Studio
Amazon SageMaker feature
SageMaker Canvas
Amazon SageMaker feature
Notebooks
IoT Analytics feature
Autopilot
Amazon SageMaker feature
Blogs
See all 672 results
Rightsizing Amazon SageMaker endpoints
By: Durga Sury and Victor Jaramillo | Date: May 19, 2022
Detect adversarial inputs using Amazon SageMaker Model Monitor and Amazon SageMaker Debugger
By: Nathalie Rauschmayr, Sergul Aydore, Yigitcan Kaya, Bilal Zafar | Date: April 5, 2022
Hierarchical Forecasting using Amazon SageMaker
By: Mani Khanuja, Farooq Sabir, Neha Gupta | Date: December 10, 2021
Announcing Amazon SageMaker Inference Recommender
By: Sean M. Tracey | Date: December 1, 2021
Documentation
See all 63,071 results in Documentation
SageMaker Roles - Amazon SageMaker
Developer Guide
Feedback
Looking for language selection? Find it in the new Unified Settings
© 2022, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences

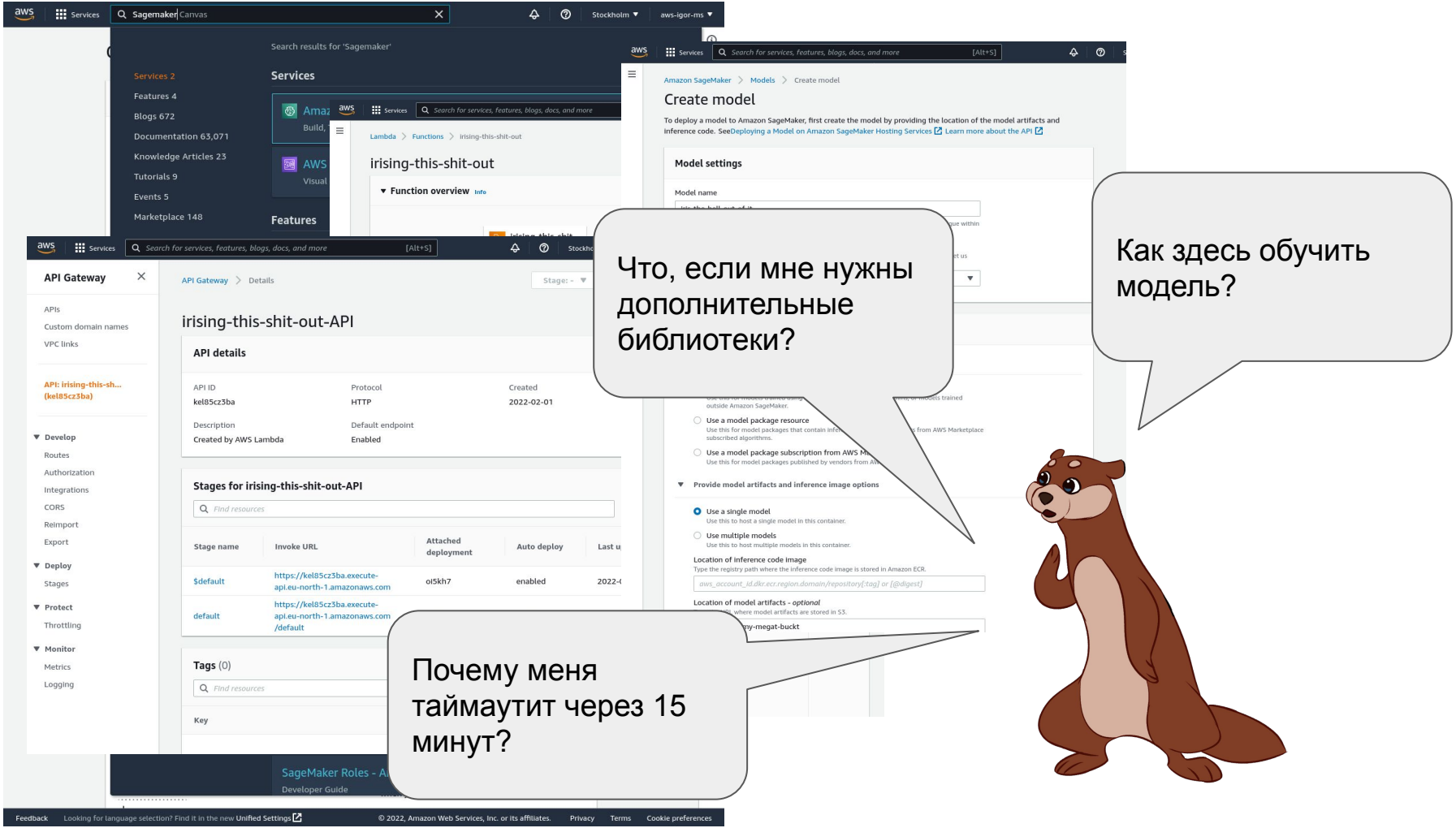

О, Бен же говорил
про Sagemaker



И про Lambda, кажется



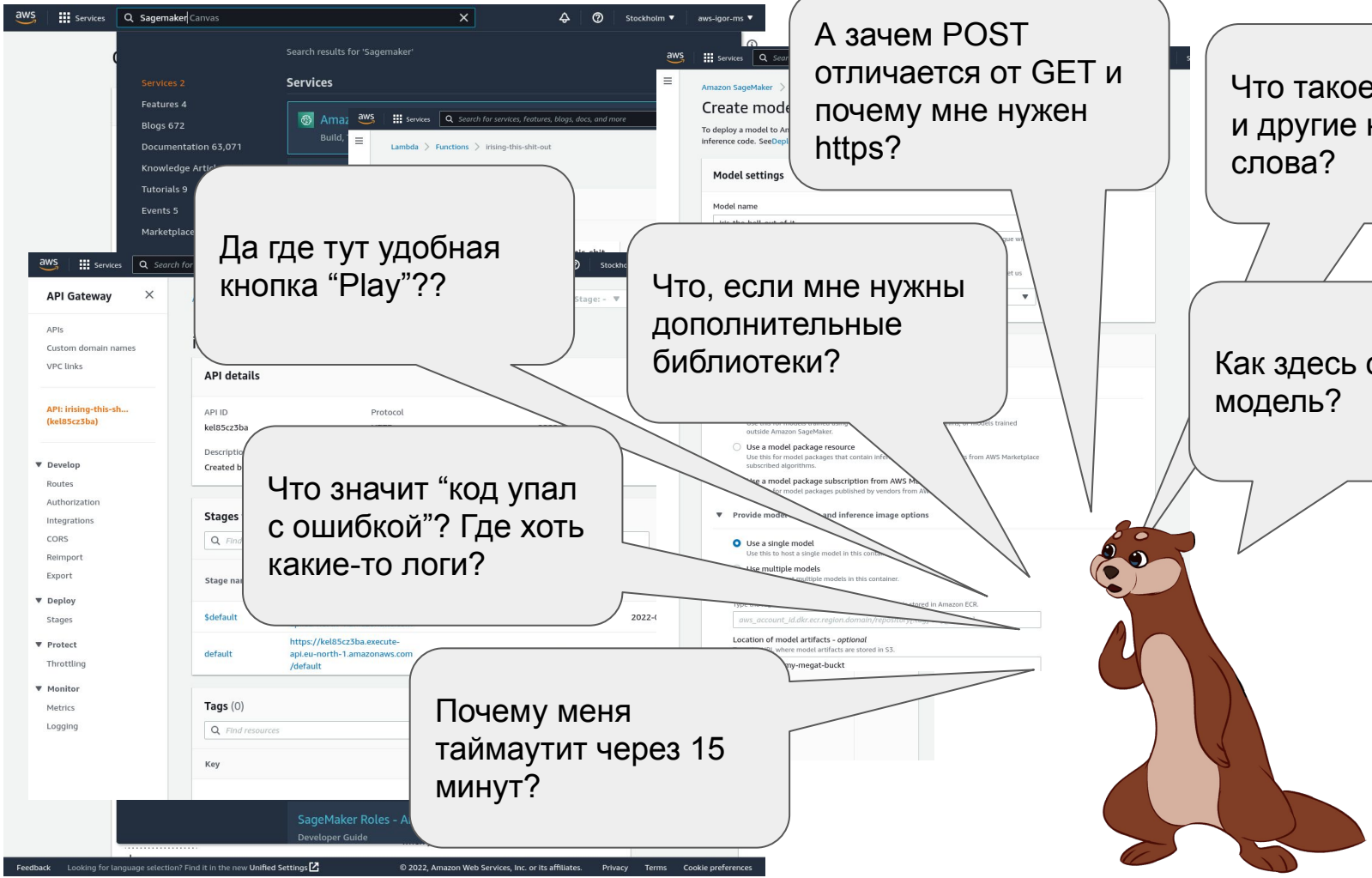
…ещё был какой-то gateway

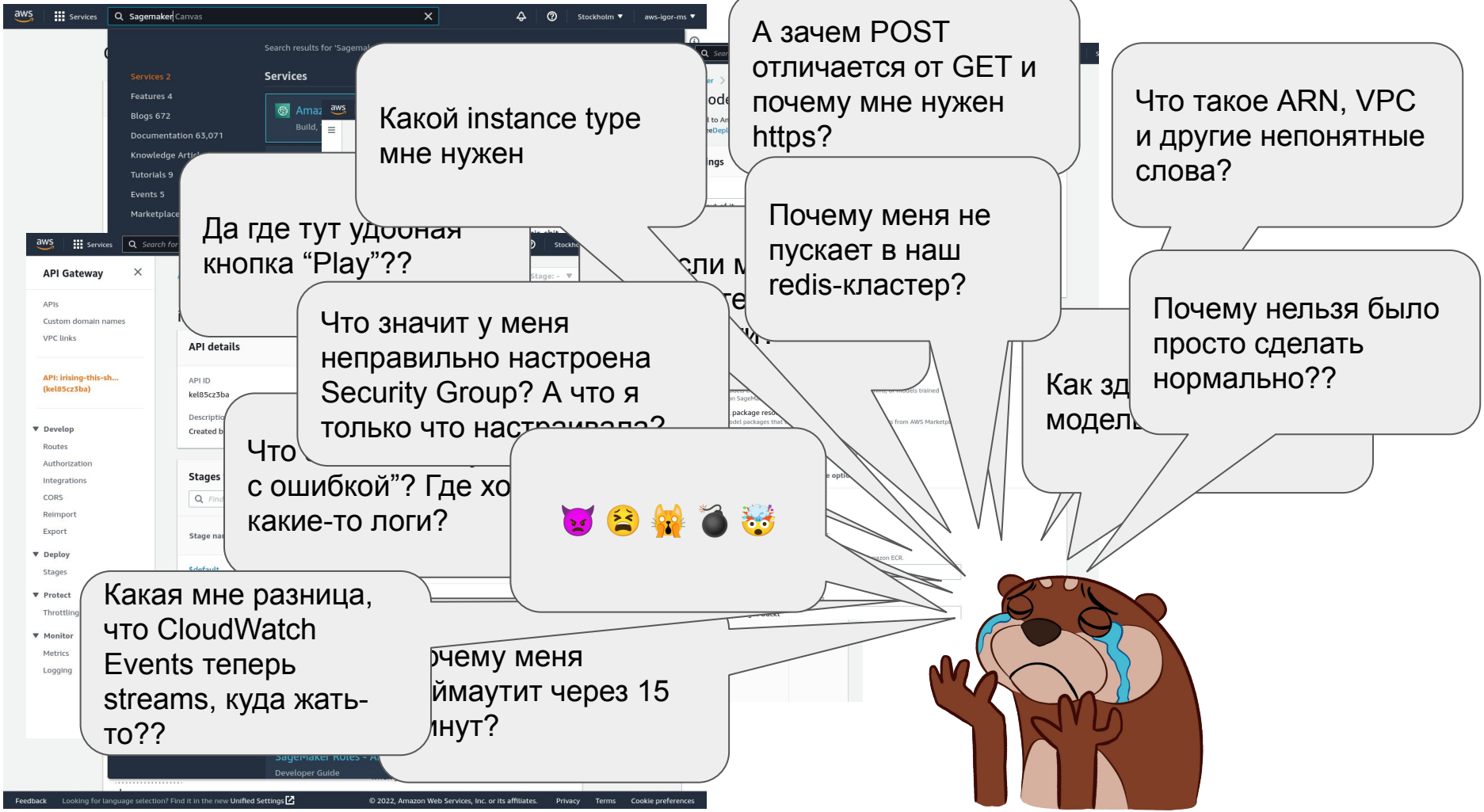


Что, если мне нужны дополнительные библиотеки?

Как здесь обучить модель?

Почему меня таймаутит через 15 минут?


Да где тут удобная кнопка “Play”??
Что, если мне нужны дополнительные библиотеки?
А зачем POST отличается от GET и почему мне нужен https?
Что такое ARN, VPC и другие непонятные слова?
Как здесь обучить модель?
Что значит “код упал с ошибкой”? Где хоть какие-то логи?
Почему меня таймаутит через 15 минут?


Какой instance type мне нужен
А зачем POST отличается от GET и почему мне нужен https?
Что такое ARN, VPC и другие непонятные слова?
Да где тут удобная кнопка “Play”??
Почему меня не пускает в наш redis-кластер?
Почему нельзя было просто сделать нормально??
Что значит у меня неправильно настроена Security Group? А что я только что настраивала?
Что с ошибкой”? Где хо какие-то логи?
😈 😫 🙀 💣 🤯
Какая мне разница, что CloudWatch Events теперь streams, куда жать-то??
очему меня ймаутит через 15 инут?
Как зд модел
API Gateway
APIs
Custom domain names
VPC links
Develop
Routes
Authorization
Integrations
CORS
Reimport
Export
Deploy
Stages
Protect
Throttling
Monitor
Metrics
Logging
API details
Stages
Services 2
Features 4
Blogs 672
Documentation 63,071
Tutorials 9
Events 5
Marketplace
Developer Guide
Feedback
© 2022, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences

ПОЧЕМУ ТАК СЛОЖНО??

Какой instance type мне нужен

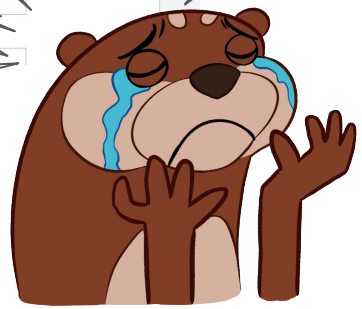
А зачем POST отличается от GET и какой мне нужен

Что такое ARN, VPC и другие непонятные слова?

Почему нельзя было просто сделать нормально??

Да где тут удобная кнопка “Play”??

Какая мне разница, что CloudWatch Events теперь streams, куда жать-то??

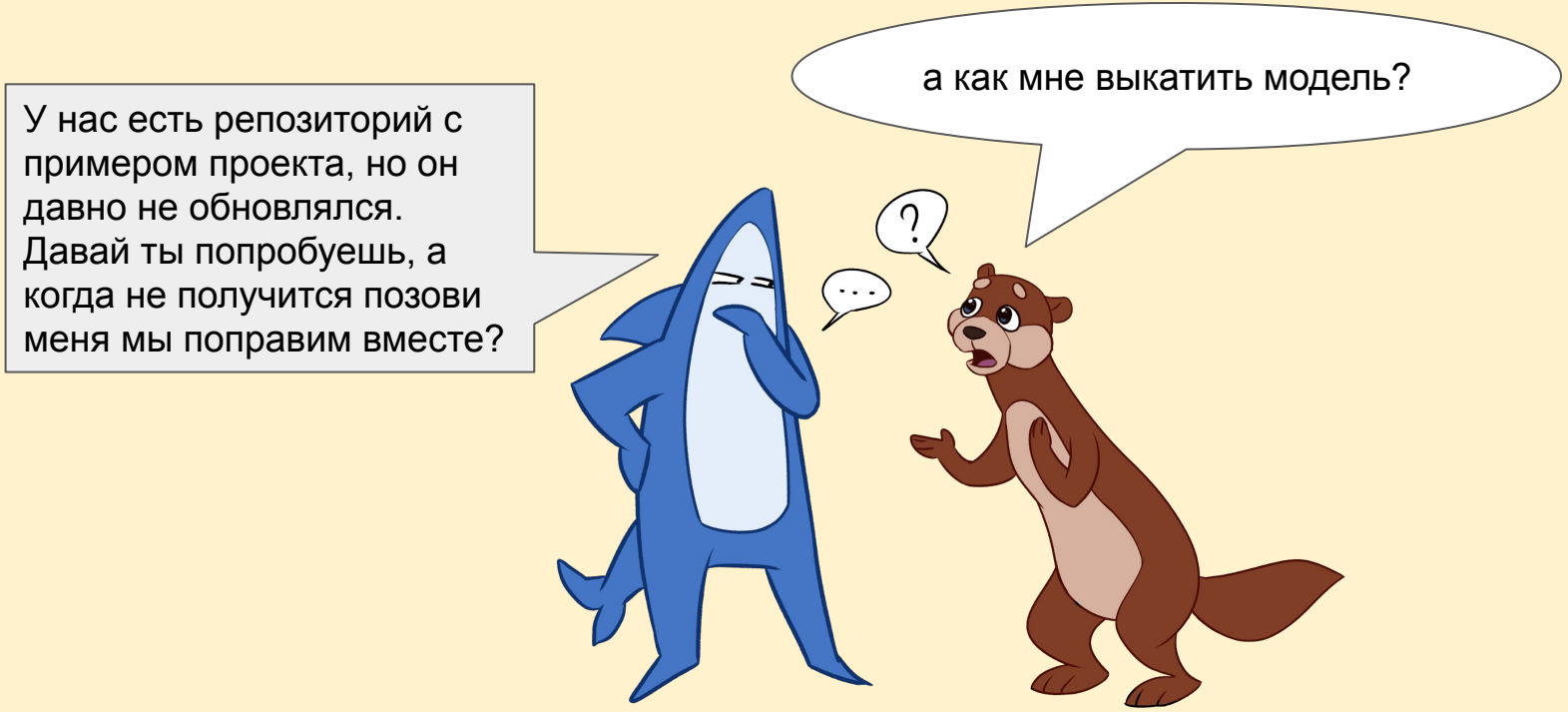


иллюзия лёгкости ломается очень быстро

- все туториалы и документация рассматривают что-то простое
- любой шаг влево-вправо карается непониманием
- лавина непривычных терминов, даже если ты не первый раз тут
- а как понять, сколько это стоит – вообще отдельная история
- если вы думаете, что так с AWS, а на каком-нибудь другом сервисе будет попроще, то увы нет (потому что проблема не в конкретном облаке)

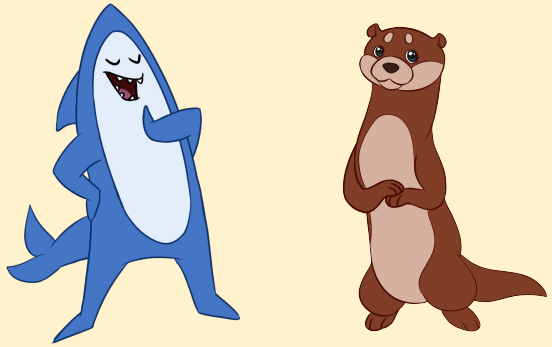
как этого избежать?

представьте:
инженер помогает аналитику, но делает это с уважением



как можно этого добиться и что делать

- документация важна, и пусть на неё будет время
- демо и бахвальство-хвастовство — ЭТО НОРМАЛЬНО
- всякие внутренние минихакатоны* тоже полезно проводить
- инженеры могут настроить CI/CD и показать, как писать тесты
- базовые шаблонные проекты под типичные задачи помогут ВСЕМ
- всячески искореняйте мышление “мы-они”



*хакатоны сложно проводить нормально чтобы никто не устал, делайте это аккуратно

общие советы

- посмотрите на свои процессы
- может, стоит посидеть с кем-то из своих “кастомеров” — как они ведут процесс (знают ли они как сделать пулл-реквест там, где вы это просите?)
- Вячеслав-driven development это плохо, но хотя бы дайте Вячеславу писать внутренний ньюслеттер
- повторяющиеся вопросы? Прекрасный повод обновить документацию
- вообще, добавить аналитику просмотров в ваш портал документации — это очень полезная и простая в реализации идея
- регулярно проверяйте, что у аналитика нет ощущения, что он один

мини-интерлюдия

поднимите руку те, кто узнал коробку

или как минимум есть догадку, откуда она

поддержка героя

«Перед заключенными в камере лежит коробка с пирожными. Коробку прислали почтой из соседней диктаторской республики. А значит, на посылке должны быть марки. Но что изображено на марках? Конечно, портрет диктатора. Рисуем портрет, уменьшаем и печатаем на марках. Если посылка была отправлена, поверх марок должны быть штампы? Делаем дизайн, гравируем, ставим штампы. Коробка повернута к камере под таким углом, под которым зритель их увидеть не может. И это не перестраховка — крупного кадра не предполагается в принципе.»

подробнее: https://www.youtube.com/watch?v=-MK5VGBvOho
видео, где это описывает дизайнер фильма

сама цитита-пересказ из телеграм-канала “плавучая редакция” @editboat

один супер-конкретный совет-демонстрация 🐩🦭🍕

```python
dictionary = {
    "word": 4,
    "bork": 10,
    "nice": 5,
}
print(dictionary.items())

dict_srtd_by_key =  dict(sorted(dictionary.items()))
print(dict_srtd_by_key)

count_getter = lambda item: item[1]

dict_srtd_by_value = dict(sorted(dictionary.items(), key=count_getter))
print(dict_srtd_by_value)

```

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL GITLENS JUPYTER

```
± |master ?:1 ✗| → cd /home/igor.mosyagin/_PC/intro-course ; /usr/bin/env /home
/igor.mosyagin/.pyenv/versions/3.10.2/bin/python /home/igor.mosyagin/.vscode/ext
ensions/ms-python.python-2022.8.0/pythonFiles/lib/python/debugpy/launcher 33471
-- /home/igor.mosyagin/_PC/intro-course/part3-word-counter/examples.py
dict_items([('word', 4), ('bork', 10), ('nice', 5)])
{'bork': 10, 'nice': 5, 'word': 4}
{'word': 4, 'nice': 5, 'bork': 10}
```

```python
dictionary = {
    "word": 4,
    "bork": 10,
    "nice": 5,
}
print("👉", dictionary.items())

dict_srtd_by_key =  dict(sorted(dictionary.items()))
print("🤔", dict_srtd_by_key)

count_getter = lambda item: item[1]

dict_srtd_by_value = dict(sorted(dictionary.items(), key=count_getter))
print("🙌", dict_srtd_by_value)

```

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL GITLENS JUPYTER

```
± |master ?:1 ✗| → /usr/bin/env /home/igor.mosyagin/.pyenv/versions/3.10.2/bin/
python /home/igor.mosyagin/.vscode/extensions/ms-python.python-2022.8.0/pythonFi
les/lib/python/debugpy/launcher 38645 -- /home/igor.mosyagin/_PC/intro-course/pa
rt3-word-counter/examples.py
👉 dict_items([('word', 4), ('bork', 10), ('nice', 5)])
🤔 {'bork': 10, 'nice': 5, 'word': 4}
🙌 {'word': 4, 'nice': 5, 'bork': 10}
```

ИТОГО

Тезисы-выводы етц

- Аналитикам: будьте более инженерными, если получается
- Инженерам: будьте более коммуникабельные, покажите аналитикам, как им жить
- Менеджерам: узнавать как у всех дела, помогать фасилитировать
- Не надо пытаться стать друг другом, симбиоз — это круто!
- Взаимное обучение (друг у друга) помогает жить примерно всем
- Пишите инструменты друг для друга в том числе, не просто же тикеты по борде таскать
- **Давайте жить дружно**

Get in touch

- https://newpodcast2.live ламповый подкаст про айти-темы
- https://debrief.site дата-инженерный дайджест на ломаном английском
- соцсети: @shrimpsizemoose везде, где найдёте

если вам тоже нужны картинки
https://linktr.ee/draktau