

The Plesk logo is displayed in white lowercase letters on a dark blue background. The letter 'p' has a small white underline. The background of the slide features a decorative pattern of overlapping triangles in various shades of blue and teal.

plesk

# Как команде R&D пережить Data-хайп?

Дмитрий Соловьев  
Program Manager/Data Scientist

# Цитаты великих (Герман Греф, 2017)

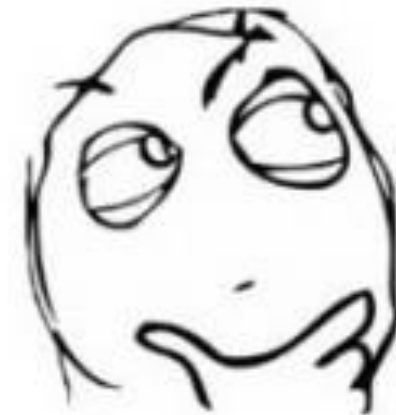
“Традиционные IT-технологии уже уходят в прошлое, а их место занимают Big Data”

“Сегодня все, кто не использует Big Data — это учреждения вчерашнего дня”

“Назовите мне типы нейронных сетей? Не знаете, двоечники! Хочу вам сказать, что это недопустимо. Вы — студенты вчерашнего дня.”

# Цитаты великих

“Data is the new oil.”



# Plesk

- Почти 20 лет развития
- 6% интернета
- 380 000 серверов
- 11 000 000 сайтов
- Коробочный продукт



# Эпизод 1 – Скрытая угроза

1. Изначально R&D далёк от data science.
2. Мы должны стать data-driven company!
3. R&D должен в зоне своей ответственности всё “оцифровать” и перейти на data-driven decisions.
4. ???



# Data Science

- Поддержка принятия решений людьми
- Автоматическое принятие решений
- Всякая всячина:
  - классификация изображений
  - подделка видео
  - генерация текстов
  - и так далее

# Не проект, но продукт

Внутренние проекты в R&D не всегда результатом имеют продукт (“клиенты” слишком в теме).

Внутренний data science проект должен выдавать продукты, то есть:

- готовые к употреблению
- документированные
- сопровождающиеся обучением и поддержкой пользователей

# Типовым задачам – типовые решения

Если ваша задача стандартна и полностью решается подключением готового чужого продукта – пользуйтесь, счастливики!

Касается не только продукта целиком.





# Задача типовая, но не совсем?

Если использовать готовый продукт как часть решения – сможете ли вы в дальнейшем соединить данные?

Особенно в исторической перспективе.

Особенно для SaaS-инструментов.



# Делаем сами – какими силами?

Обычная команда разработки (часть существующей?)

+

Data Scientist:

- понимающий предметную область
- знающий статистику и тервер
- умеющий программировать
- ...



# Data Engineer – кто такой?

DevOps для данных.

Data Engineer готовит инфраструктуру и доставку данных, Data Scientist работает с данными.

Но слишком хорошо – тоже нехорошо. Изоляция Data Scientist'а от процесса доставки и подготовки данных опасна.

# Что в R&D “растёт само”?

Инженерам требуется:

- принципиальное понимание: да/нет, используется мало/заметно/много и т.п.
- не постоянно, а когда требуется для принятия решения
- информация, которую можно понять – даже если придётся подумать

# Взгляд с другой стороны

Топы воспринимают иначе:

- требуется (привычна) бухгалтерская точность
- в динамике и в любой момент времени
- в лёгкой для восприятия форме



# Точность, надёжность и актуальность

Если собираете данные, которые не 100% точны – прячьте! Но это ненадолго – данные просачиваются.

Необходима инвентаризация данных – что, откуда, как, кто ответственный, качество данных, возможность связать их с остальными?



# Как обеспечить?

Сбор и обработка данных – такая же часть функционала продукта, как и прочие.

От определения требований – до регрессионного тестирования и поддержки.



# Данные не стареют

Храните историю, предусмотрите её доступность. Логи – хорошо, но недостаточно. Идеально – возможность отката и повторной обработки.

Воспроизведение состояния набора данных на определённый момент в прошлом.

При выборе внешнего решения – обратите внимание.



# (Без)опасность данных

Заботьтесь о безопасности, конфиденциальности и т.д.

Если работаете с Европой – изучите GDPR.

Инвентаризация данных!



# Метрики

Если есть какие-нибудь числа, связанные с продуктом, то из них наверняка сделают метрики. Особенно если числа растут со временем.

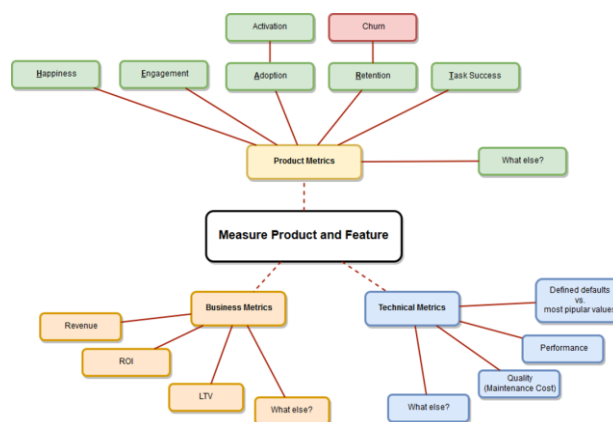
Не давайте бесконтрольно плодить метрики! KPI с прямым влиянием на P&L – надёжней всего.

Метрики (конфигурация, вопросы) как код. Code review, система управления версиями и т.д.

# Роль данных в процессе разработки

“Покажите мне \$\$\$” – радикальное решение. Самое убедительное, но очень негибкое.

С самого начала (с требований) определяем наши гипотезы, чуть позже – метрики успеха. Какие можем.



# А как же ML, AI, etc?

80% работы – подготовка данных.

Всё уже украдено до нас – ищите готовые решения аналогичных задач.

Если можно сделать без ML – делайте без.

**Забудьте о моде, не ставьте технологии вперёд – трезво оценивайте себя и свои задачи.**

# Вопросы?

Дмитрий Соловьев

Program Manager/Data Scientist

✉ [dsolovyev@plesk.com](mailto:dsolovyev@plesk.com)

f [dmitry.solovyev.1977](https://www.facebook.com/dmitry.solovyev.1977)

plesk



Присоединяйся!

[career@plesk.com](mailto:career@plesk.com)