

Рекомендательные сервисы в продакшене

Николай Анохин

22 февраля 2023 г.

Архитектуры рекомендательных сервисов
●oooooooooooo

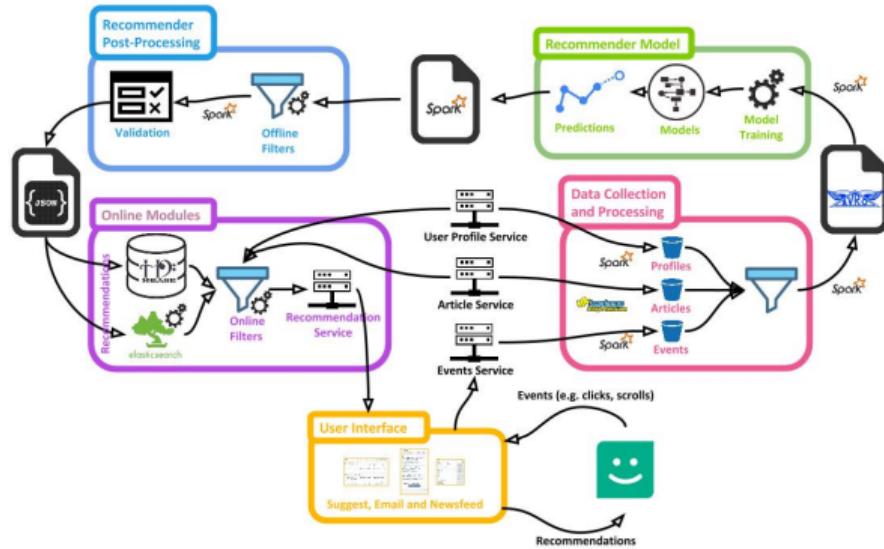
Метрики и эксперименты
oooooooooooo

Итоги
оооо

Архитектуры рекомендательных сервисов



Обзор типичных компонентов RS / Mendeley (2016) [JIH16]



Машинное обучение – небольшая часть рекомендательного сервиса. Другие компоненты часто требуют не меньше усилий.

Почта Чат и чат-группы От Любой время Содержит прикрепленные файлы Кому Не прочитано Расшире > 1–50 из множества < > Ры

□ ★ Mendeley Входящие "Values of user exploration in recommender systems..." and more articles on Mendeley - ... Mendeley 9 сент.

□ ★ Mendeley "Values of user exploration in recommender systems..." and more articles on Mendeley - ... Mendeley 3 сент.

□ ★ Mendeley Входящие "Values of user exploration in recommender systems..." and more articles on Mendeley - ... Mendeley 26 авг.

□ ★ Mendeley Входящие "Values of user exploration in recommender systems..." and more articles on Mendeley - ... Mendeley 19 авг.

□ ★ Mendeley Входящие "Values of user exploration in recommender systems..." and more articles on Mendeley - ... Mendeley 12 авг.

□ ★ Mendeley Входящие "Values of user exploration in recommender systems..." and more articles on Mendeley - ... Mendeley 5 авг.

□ ★ Mendeley Входящие "Values of user exploration in recommender systems..." and more articles on Mendeley - ... Mendeley 29 июл.

□ ★ Mendeley Входящие "Values of user exploration in recommender systems..." and more articles on Mendeley - ... Mendeley 22 июл.

□ ★ Mendeley Входящие "Values of user exploration in recommender systems..." and more articles on Mendeley - ... Mendeley 8 июл.

□ ★ Mendeley Входящие "Values of user exploration in recommender systems..." and more articles on Mendeley - ... Mendeley 2 июл.

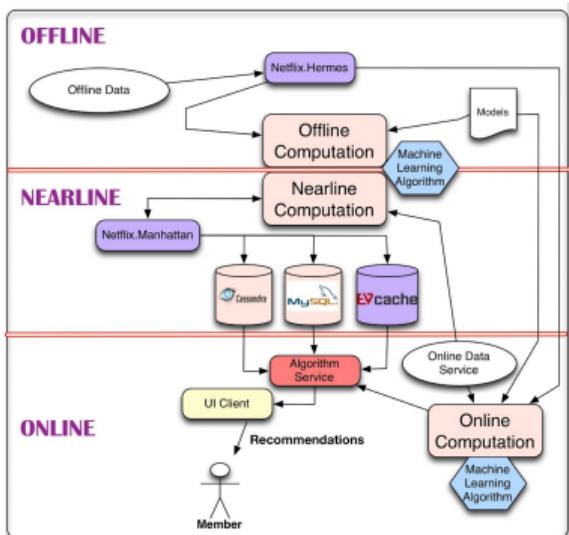
□ ★ Mendeley Входящие "Values of user exploration in recommender systems..." and more articles on Mendeley - ... Mendeley 24 июн.

□ ★ Mendeley Входящие "Values of user exploration in recommender systems..." and more articles on Mendeley - ... Mendeley 17 июн.

Показать подробные сведения



Обработка данных под высокой нагрузкой / Netflix (2013) [NN13]

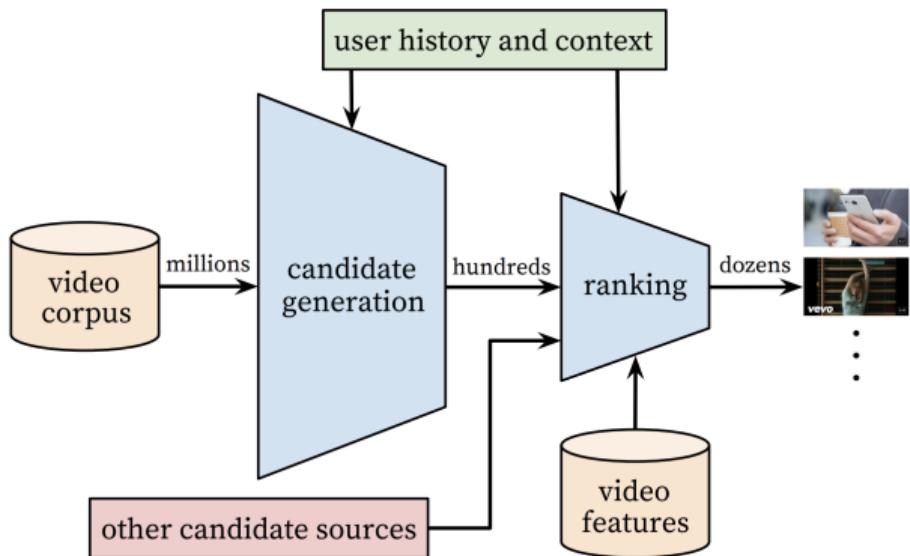


Двигаясь от offline к real-time, мы можем быстрее реагировать на изменения контекста. При этом возникают ограничения на сложность алгоритмов.

The screenshot shows the Netflix homepage with a dark background. At the top, there's a banner for the TV show "THE TINDER SWINDLER" with a "1" badge. Below it are buttons for "Play" and "More Info". To the right, there are two orange circular icons with the numbers "6" and "5", likely indicating the count of new or popular shows. The main content area features a grid of TV show thumbnails. In the "TV Shows" section, shows like "THE WOMAN IN THE HOUSE", "OZARK", "THE SINNER", "STAY CLOSE", "After Life", and "DOWNTON ABBEY" are displayed, each with a "TOP 10" badge and a "NEW EPISODES" button. Below this, the "Trending Now" section shows thumbnails for "Sweet Magnolias", "THE TINDER SWINDLER" (with a "3" badge), "THE POWER OF THE DOG", "THE CROWN", "BRIDGERTON", and "MONEY HEIST". Each thumbnail includes a "TOP 10" badge and a "NEW EPISODES" button. The Netflix logo is visible in the bottom right corner.



Рекомендации айтемов из больших каталогов / Youtube (2016) [CAS16]



Айтемов так много, что учесть **полный контекст** не может даже Google. Для быстрого отбора кандидатов применяются грубые фильтры.

Загадка

Что общего между

- населением городов
- количеством друзей у пользователей в социальной сети
- размерами лесных массивов
- количеством прослушиваний песен в Spotify



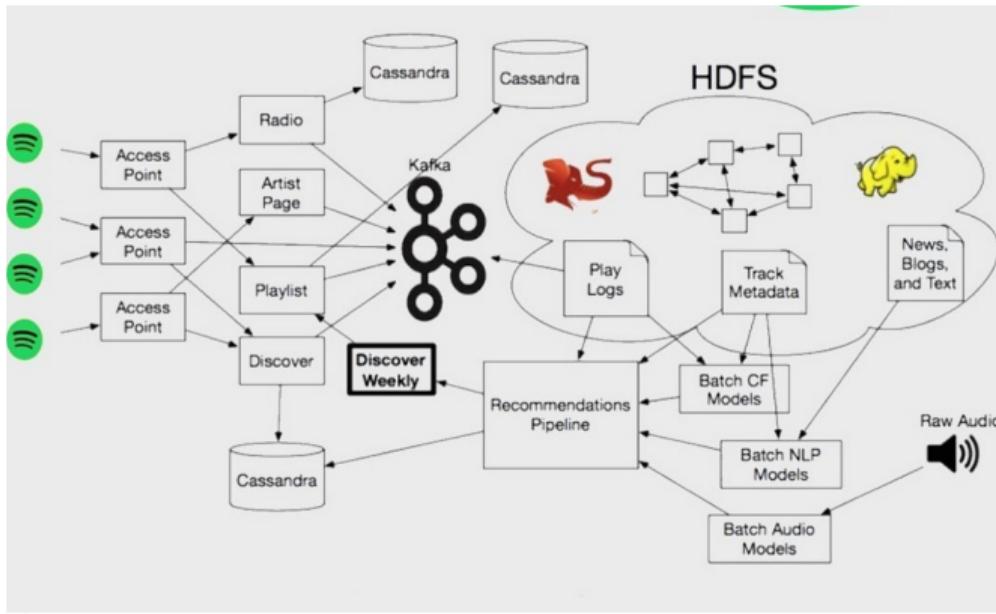
Power law

$$p(x) = \frac{C}{x^\alpha}, \quad x > x_{min}$$



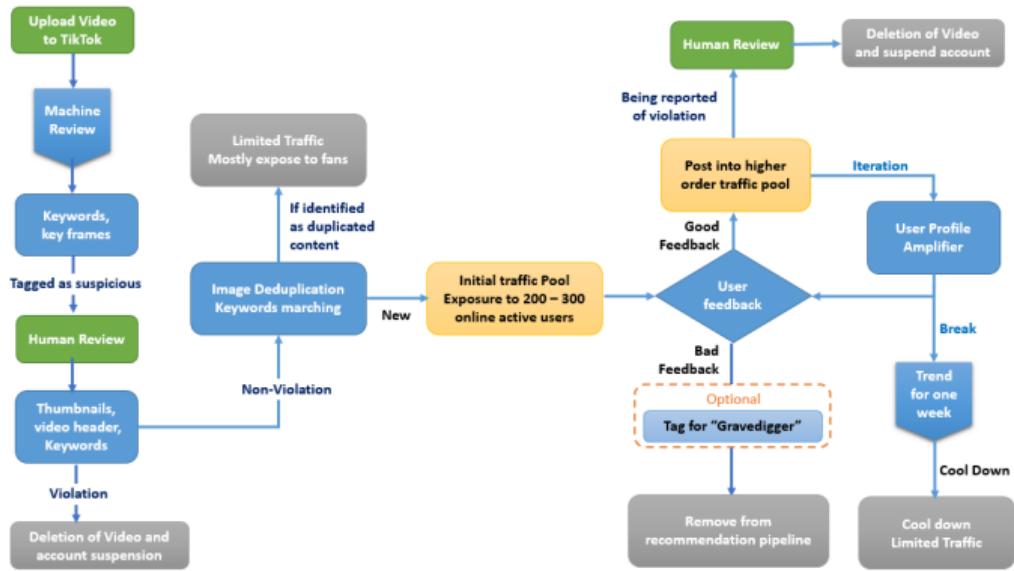
Правило 80/20

Холодный старт и длинный хвост / Spotify (2016) [Spo16]



Холодные айтемы и пользователи будут всегда. Использование контента - один из вариантов решения проблемы

Несоответствие таргетов моделей и business-value / TikTok (2020) [Wan20]



Потребности людей нельзя упаковать в удобную метрику. Кроме машинного обучения в рекомендательных сервисах приходится использовать пре- и пост-процессинг, чтобы гарантировать business-value.

Как в действительности выглядит архитектура RS



Какие сложности учитывает архитектура RS

- Высокая нагрузка рекомендательных сервисов
- Большие каталоги айтемов
- Холодный старт пользователей и айтемов
- Несоответствие business-value и метрик оптимизации

Какие технические средства могут понадобиться

- Отказоустойчивые продакшен-сервисы (HBase, Cassandra, Elasticsearch)
- Передача данных (kafka)
- Хранение данных (Hadoop HDFS)
- Batch обработка данных (Spark)
- Потоковая обработка данных (Kafka, Spark Streaming)



Архитектуры рекомендательных сервисов
oooooooooooo

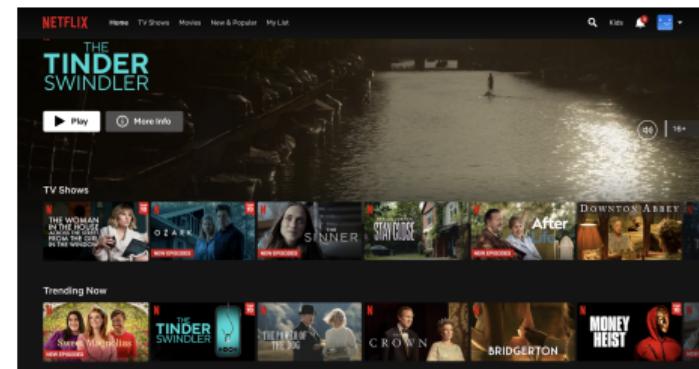
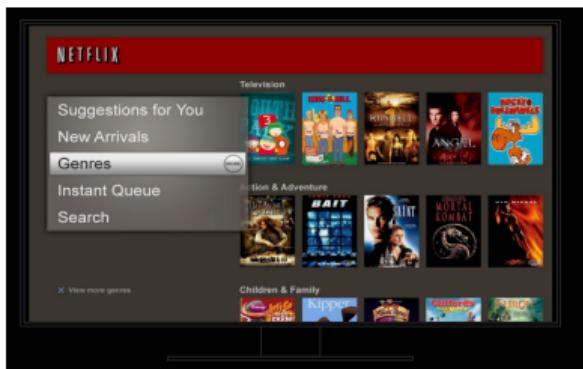
Метрики и эксперименты
●oooooooooooo

Итоги
oooo

Метрики и эксперименты



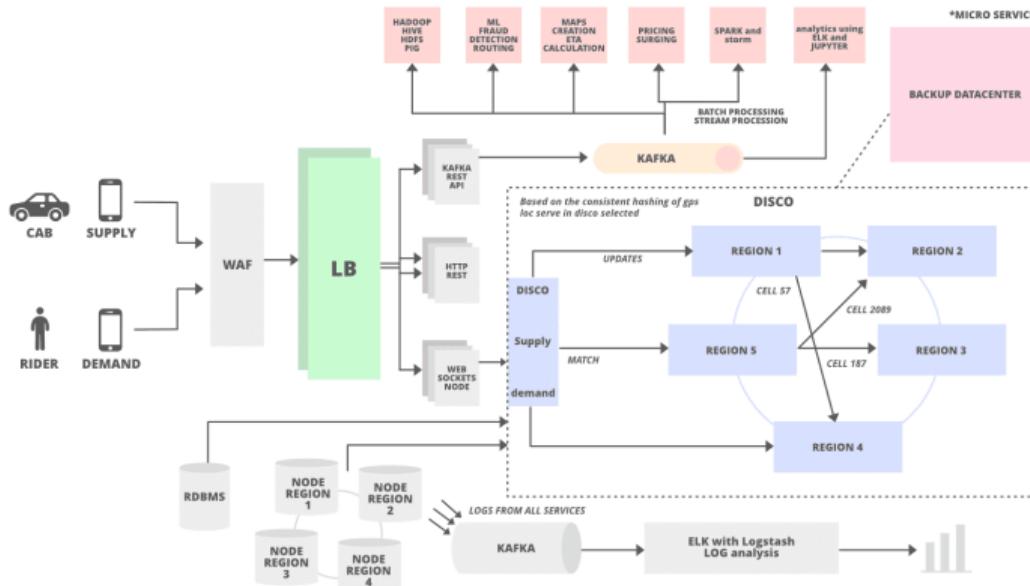
Netflix 2010-2021 [NET21]



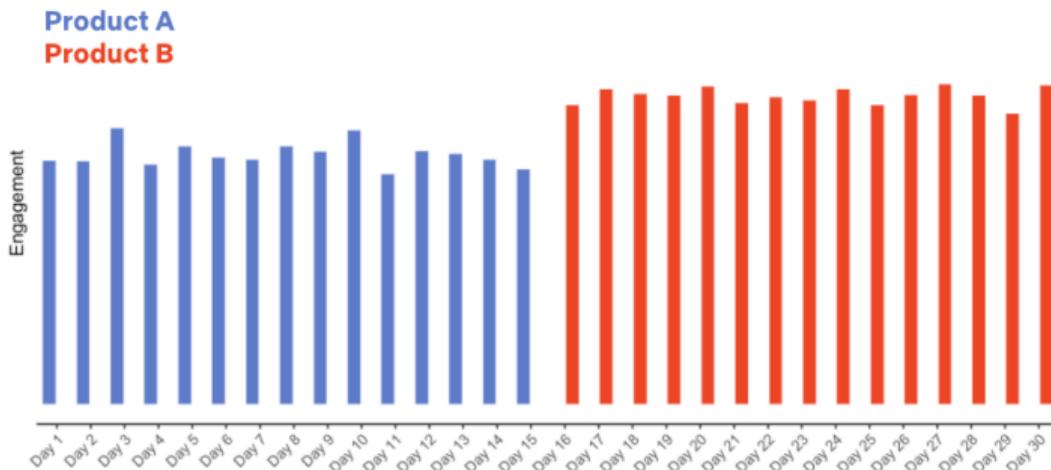
Хотим принимать решения на основе данных →

Начинаем собирать метрики →

Разрабатываем инструменты для принятия решений



Наивный подход к измерению эффекта



Задача

Какой причинно-следственный эффект на распределение целевой метрики окажет выбранное воздействие T ?

Фундаментальная Проблема Causal Inference

Для конкретного пользователя невозможно вычислить causal effect напрямую, потому что нельзя проанаблюдать значение целевой переменной при более чем одном значении T^a

^aБез дополнительных предположений эту проблему не решить [GH07]



Randomized Controlled Experiment

Схема эксперимента

Все доступные пользователи независимо друг от друга случайным образом распределяются в control либо treatment с одинаковой вероятностью

Предположение 1:

Можно оценить значение некоторой характеристики для всей популяции, имея выборку из этой популяции.

Предположение 2: Stable Unit Treatment Value Assumption

Эффект для каждого пользователя зависят только от свойств этого пользователя, но не свойств и исходов других пользователей.



Фреймворк Potential Outcomes

Воздействие на i пользователя:

$$T_i = \begin{cases} 0, & \text{если показываем control} \\ 1, & \text{если показываем treatment} \end{cases}$$

Соответствующие потенциальные исходы:

$$y_i^0 \text{ и } y_i^1$$

Требуется оценить:

Average Treatment Effect

$$ATE = E [y_i^1 - y_i^0]$$

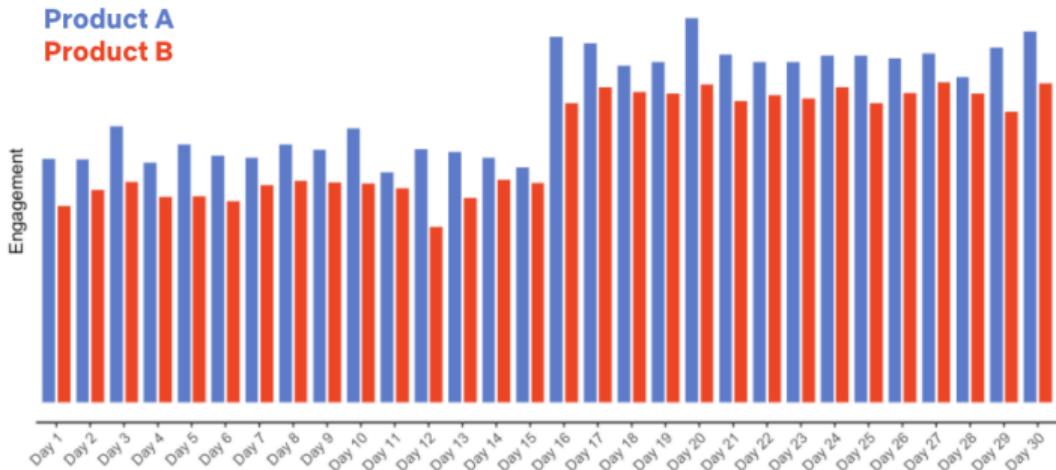


Оцениваем ATE в RCE

$$ATE = E[y_i^1 - y_i^0] = E[y_i^1] - E[y_i^0] \sim \text{avg}_{i \in T}(y_i^1) - \text{avg}_{i \in C}(y_i^0) = \bar{y}_1 - \bar{y}_0$$

- нужно оценить две характеристики – $E[y_i^0]$ и $E[y_i^1]$, поэтому используем выборки C и T
- проще всего сделать оценку, если выборка несмещенная
- чем больше данных, тем точнее оценка





Доверительный интервал на ATE

Доверительный интервал (L, U) с уровнем доверия α :

$$P(L < \theta < U) = 1 - \alpha$$

Формула Уэлча:

$$\bar{y}_1 - \bar{y}_0 \pm t_{\alpha/2,r} \sqrt{\frac{s_1^2}{n_1} + \frac{s_0^2}{n_0}}, \quad r = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_0^2}{n_0}\right)^2}{\frac{s_1^4}{n_1^2(n_1-1)} + \frac{s_0^4}{n_0^2(n_0-1)}}$$

Где:

- n_1 и n_0 – количество пользователей в treatment и control
- s_1^2 и s_0^2 – оценки дисперсии метрики в treatment и control
- $t_{\alpha/2,r}$ – табличное значение для r степеней свободы



На практике

- Перед запуском
 - Выбираем ключевую метрику, несколько сопутствующих метрик и контролируем, что не “уронили” важные
 - Выбираем длительность эксперимента, оценивая мощность теста :D
- При анализе
 - Метрики распределены по-разному: нужно подбирать подходящие тесты
 - Используются методы снижения дисперсии оценок (cuped, diff-in-diff)

Если вы попали в компанию, в которой есть культура принятия решений на основе данных – сохраняйте ее всеми силами. Если нет – пропагандируйте.



Сложности RCE в индустриальных рекомендерах [GKT⁺19]

- Как выбрать Самую Главную Метрику (OEC)?
- Как оценить долгосрочный эффект?
- Разный эффект на разных сегментах пользователей
- Отсутствие культуры экспериментирования
- Масштабирование платформы для экспериментов
- Сетевые эффекты
- Наложение эффектов от экспериментов



Архитектуры рекомендательных сервисов
oooooooooooooo

Метрики и эксперименты
oooooooooooo

Итоги
●ooo

Итоги



В основе рекомендательных сервисов лежит машинное обучение. При проектировании нужно учитывать множество дополнительных факторов, например требования к скорости обработки данных, эффект длинного хвоста и возможность холодного старта.

А/В эксперимент – надежный способ оценки эффекта от изменений в сервисе.





Литература I

-  Paul Covington, Jay Adams, and Emre Sargin, *Deep neural networks for youtube recommendations*, Proceedings of the 10th ACM Conference on Recommender Systems (New York, NY, USA), RecSys '16, Association for Computing Machinery, 2016, p. 191–198.
-  Andrew Gelman and Jennifer Hill, *Data analysis using regression and multilevel/hierarchical models*, vol. Analytical methods for social research, Cambridge University Press, New York, 2007.
-  Somit Gupta, Ronny Kohavi, Diane Tang, Ya Xu, Reid Andersen, Eytan Bakshy, Niall Cardin, Sumita Chandran, Nanyu Chen, Dominic Coey, Mike Curtis, Alex Deng, Weitao Duan, Peter Forbes, Brian Frasca, Tommy Guy, Guido W. Imbens, Guillaume Saint Jacques, Pranav Kantawala, Ilya Katsev, Moshe Katzwer, Mikael Konutgan, Elena Kunakova, Minyong Lee, MJ Lee, Joseph Liu, James McQueen, Amir Najmi, Brent Smith, Vivek Trehan, Lukas Vermeer, Toby Walker, Jeffrey



Литература II

Wong, and Igor Yashkov, *Top challenges from the first practical online controlled experiments summit*, SIGKDD Explor. Newslett. 21 (2019), no. 1, 20–35.

-  Kris Jack, Ed Ingold, and Maya Hristakeva, *Mendeley suggest architecture*, Oct 2016.
-  *Decision making at netflix (series)*, Sep 2021.
-  Xavier Amatriain Netflix and Justin Basilico Netflix, *System architectures for personalization and recommendation*, Mar 2013.
-  GALVANIZE Spotify, *Ever wonder how spotify discover weekly works? data science*, Aug 2016.
-  Catherine Wang, *Why tiktok made its user so obsessive? the ai algorithm that got you hooked.*, Jun 2020.

