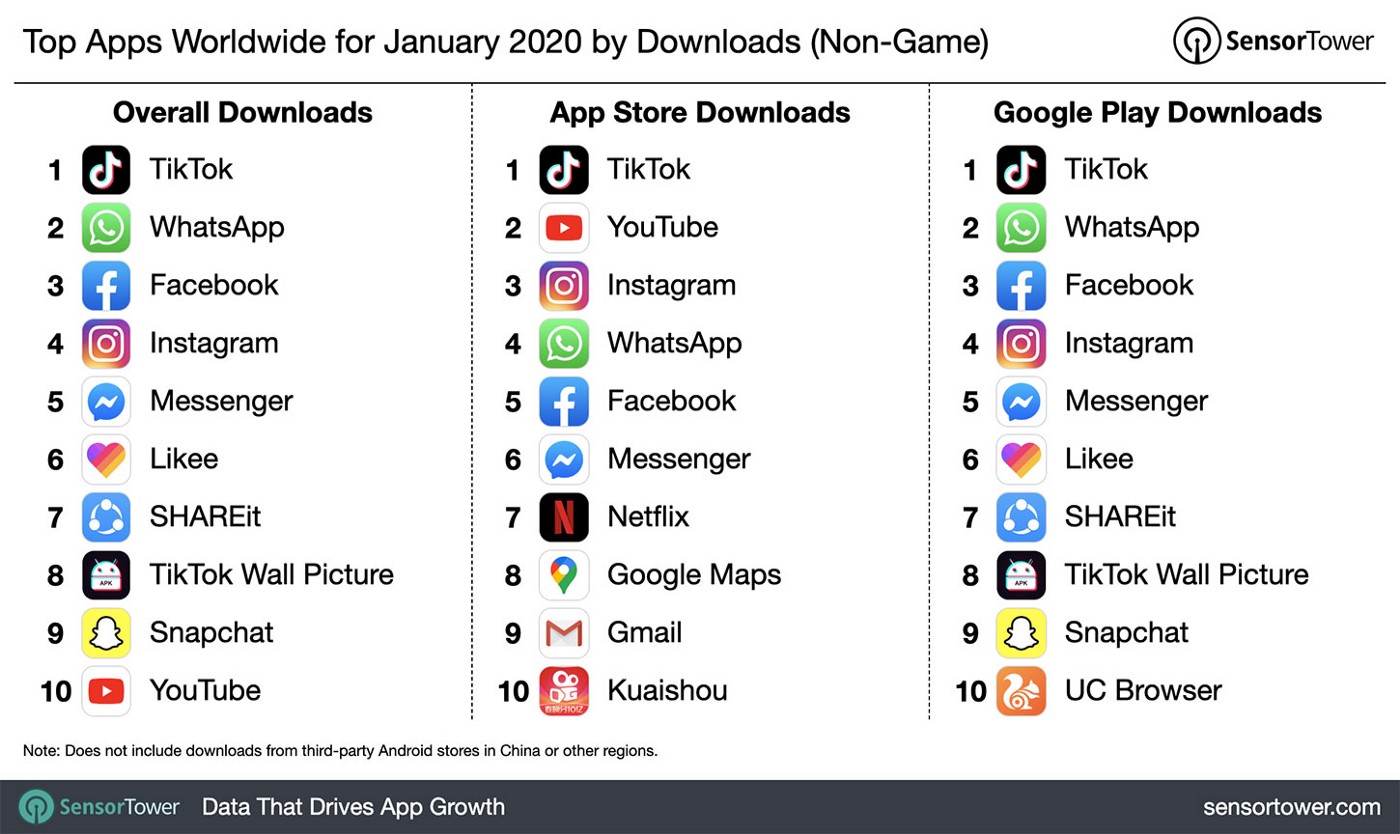
*Tick Tok is taking the world by storm. According to Sensor Tower, the short video app has been downloaded more than****2 billion times****globally on the App Store and Google Play. What’s the magic behind this sensational App that got you so obsessive? Not surprised, the answer is ML backed Recommendation Engine.*

*Table of Content (Estimated reading time 15 mins)*

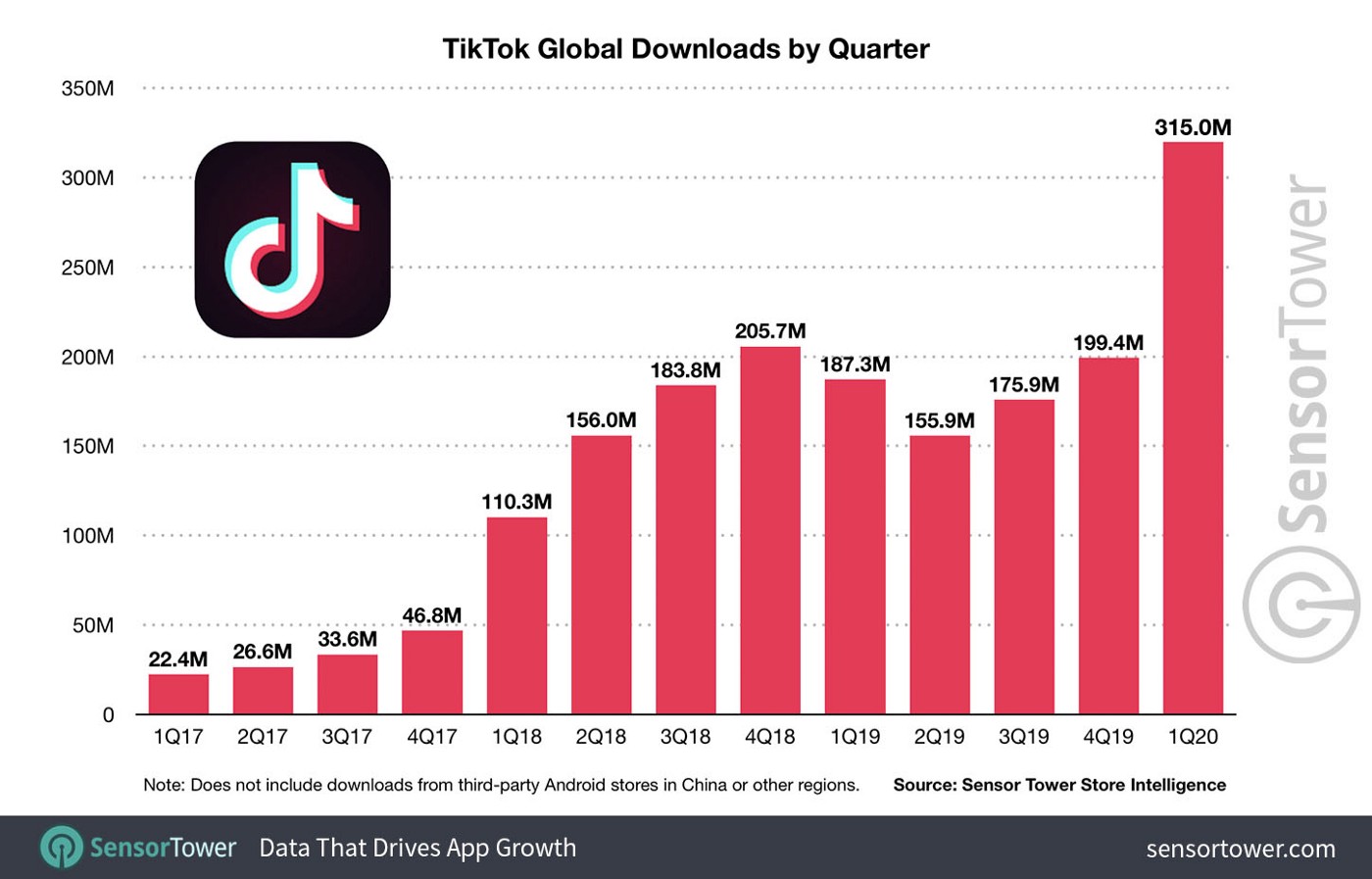
1. *General Introduction of TikTok.*
2. *The Archetype of TikTok’s Recommendation System (data, features, objectives, algorithms, and training mechanism)*
3. *TikTok’s recommendation workflow (implementation in real-time,****MUST-READ)***

OK, let’s be honest here. Who doesn’t like dog antics and funny cat videos? Especially during this grieving period of global lockdown.

But this only explains part of TikTok’s unprecedented success story. In less than 2 years, it went from a “lip-syncing” app in a small fan community to viral App with nearly 800 million active monthly users in 2020. Collectively, TikTok videos tagged with #coronavirus have been watched 53 billion times.



TikTok became the most downloaded app in the US in Jan 2020. (vis [SensorTower](https://sensortower.com/))



TikTok Total Global Downloads by [Sensor Tower](https://sensortower.com/blog/tiktok-downloads-2-billion)

It’s famous for spawning viral songs and funny mime videos.

Typically people spend 52 mins a day on the app, and this daily usage number is 26 mins, 29 mins ad 37 mins for Snapchat, Instagram, and Facebook, respectively.





Report from [Oberlo](https://au.oberlo.com/blog/tiktok-statistics)

Apart from the growth hacking strategy, this 60-second short video app is filled with memes, comedy, dancing, and talents. Equipped with one of the best[**Recommending Engine**](https://en.wikipedia.org/wiki/Recommender_system)in the industry, You don’t need to search or know whom to watch. Personalized feed was provided at a click away.

This type of endless quick stimulate of easy-to-get happiness made it hard to stop browsing on TikTok. Some people call it the ultimate time killers that suck up your spare time, and somehow create this distortion filed that “*5 mins in TikTok equals 1 hour in real life”*.

Today, we are going to discuss how did TikTok use machine learning to analyze users’ interests and preferences through the interactions then display a personalized feed for different users.

The recommendation engine is not new to the Data Science community. Instead, some consider it as the old generation AI system due to a lack of dizzying effects like image recognition or language generation.

Nevertheless, the recommendation is still one of the predominant AI systems which have the most extensive implementation in almost all online services and platforms. For example, YouTube video suggestion, campaign email you received from Amazon, book you might also like when you are browsing the kindle bookshop.

In fact, according to the [***research paper***](https://dl.acm.org/doi/10.1145/2843948)published by Gomez-Uribe and Netflix’s chief of product Neil Hunt said, the combined effect of personalization and recommendations save Netflix’s more than $1B per year. Furthermore, 80% of the subscribers are choosing videos from the engine’s suggestion list.

Now let’s take a look at what dose TikTock does differently.

1. Introduction of Recommendation Engine.

*(For those who already familiar with this topic, please jump to the next section. )*

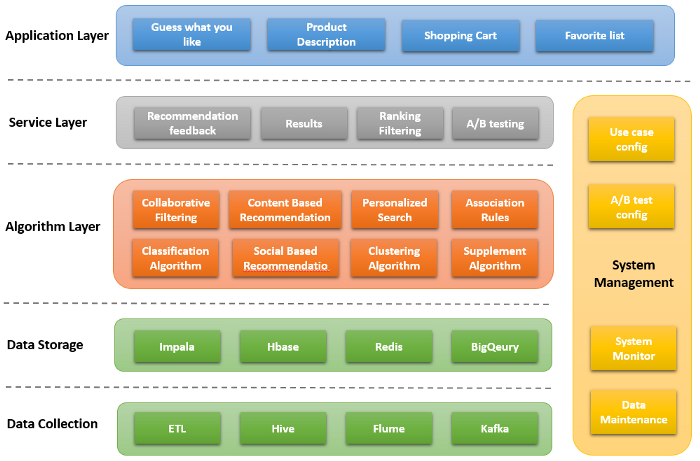
There are too many useful articles and online courses about recommendation engines, and I don’t want to re-invent the wheel.

Below are two resources for you to build up some basic knowledge for the recommendation engine.

* Comprehensive Guide to building a Recommendation Engine from scratch [[LINK](https://www.analyticsvidhya.com/blog/2018/06/comprehensive-guide-recommendation-engine-python/)] (take around 35 mins to read and 40–60 mins to replicate the python code)
* Recommendation Engine from Andrew Ng [[LINK](https://www.youtube.com/playlist?list=PL-6SiIrhTAi6x4Oq28s7yy94ubLzVXabj)] (take an hour to watch the videos)

Apart from the basic, industrialized recommendation engine need a robust backend and architecture design for integration. Below is a primary example.





Recommendation Engine — Created by Catherine Wang, All Rights Reserved

A real-time system should have a solid data basis (for collection and storage) to support the multiple abstract layers(algorithm layer, serving layer, and application layer ) on top that addresses different business problems.

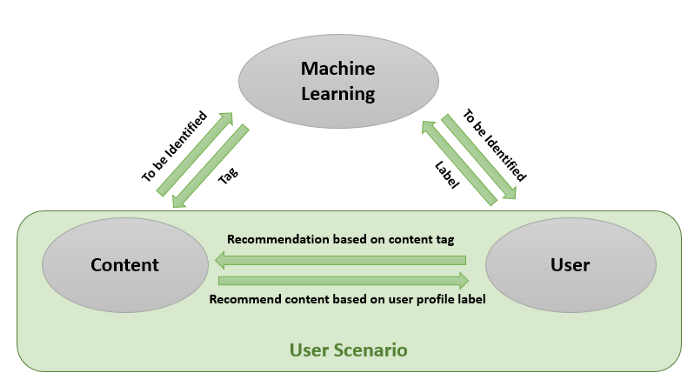
2. The Archetype of TikTok’s Recommendation System Design

‘ *User-Centric Design*’ remains the core of the archetype. In a simple term, TikTok will only recommend the content you would love, from a cold start adjustment to an explicit recommendation for active users.

If you click a dancing video, your feed would be customized to the entertainment category initially, then the following up mechanism will trace your behaviors for further analysis, which would eventually provide precise recommendations for you only.

The high-level workflow.





3 Main Components — Created by Catherine Wang, All Rights Reserved

In TikTok’s archetype, there are three main building blocks, 1) Tagging the content, 2) Creating user-profiles and user scenarios, and 3) Training and serving recommendation algorithms.

We will be discussing each of them in the following content.

2.1 Data and Features

First of all, Data. If we formally described the recommendation model, it is a function that fits a user’s satisfaction with the User Generated Content. To provide this function requires the input of data from***three-dimensions****.*





***Content Data***— TikTok is a platform with massive user-generated content. Each type of content has its traits, the system should be able to identify and distinguish them for a reliable recommendation.

***User Data****—*These include interest labels, career, age, gender, demographics, etc. It also includes latent features from ML-based customer clustering.

***Scenario data*** — This data tracks the use scenario and user’s preference shift based on different scenarios. For example, what type of video a user like to watch when they are at work, travel or commuting

Once relative data has been collected, four types of critical engineered features would be derived and feed-in into the recommendation engine.

* ***Correlation Features****:* they represent the correlation between content attributes and user tags, including keywords matching, classification tag, source matching, theme tag, and latent features like vector distances between user and content.
* ***User-Scenario Features***: engineered from scenario data including geographic location, time of the day, event tags, etc.
* ***Trend Features:*** they based on user interactions and represent a global trend, hot topic, top keywords, trend themes, etc.
* ***Collaborative Features:*** based on collaborative filtering technique. It balance the narrow recommendation (bias) and collaboration recommendation(generalization). More precisely, it will not only consider a single user’s history but analyst the collaborative behaviors of a similar user group(clicks, interests, keywords, themes).

The model will predict whether the content is suitable for the user in a scenario by learning from the above features.

2.2 Intangible objectives

In the recommended model, click-through rate, reading time, likes, comments, and reposts are all quantifiable objectives. You can use the model or algorithms to fit them then make the prediction conclusively.

However, other intangible objectives cannot be evaluated by those quantifiable indicators.

For example, to maintain a healthy community and eco-system, TikTok is aiming to suppress content that involves violence, scamming, pornography, flatulence, and weighing in facts, high-quality content like news.

For this goal, a border control frame needs to be defined beyond quantifiable model objectives. (Content Audit system)

2.3 Algorithms

The recommendation objectives can be formulated into a classic machine learning problem. Then be solved by algorithms including [collaborative filtering model,](https://en.wikipedia.org/wiki/Collaborative_filtering) Logistic Regression model, [factorization Machine,](https://en.wikipedia.org/wiki/Matrix_factorization_(recommender_systems)) GBD, and deep learning.

Картина, която съдържа крик

Описанието е генерирано автоматично

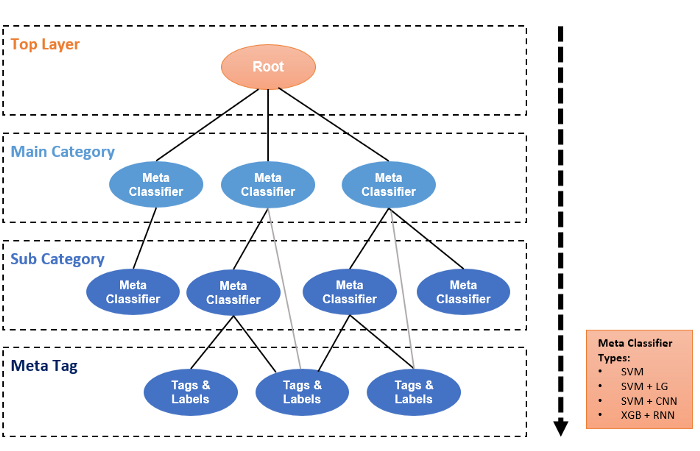


An illustration of collaborative filtering

An industrial-grade recommendation system requires a flexible and extendable ML platform to build up the experimental pipeline to train various models quickly. Then stack them to serve in the real-time. (eg. combine LR and DNN, SVM with CNN)

Apart from the main recommendation algorithm, TikTok also needs to train the content classification algorithm and the user profiling algorithm. Below is a hierarchy classification architecture for content analysis.





Hierarchical Classification Tree — Created by Catherine Wang, All Rights Reserved

Drill down from the master root. Each layer down is the main category and sub-category. Compared with a separate classifier, using a hierarchical classification mechanism can better solve the problem of data skew.

2.4 Training Mechanism

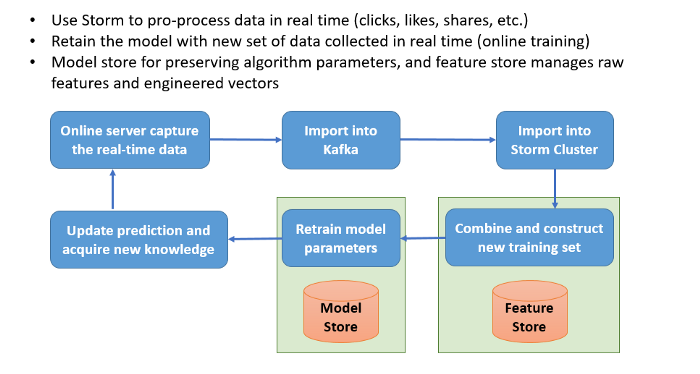
TikTok uses real-time online training protocol, it demands less computational resource and provides fast feedback. Those are important for streaming and information flow products.

The user behaviors and actions can be captured instantly, then feedback to the model to reflect on the next feed. (eg. when you click a new video, your feed will quickly change based on your latest actions)

Most likely, TikTok is using Storm Cluster to process the real-time sample data, including clicks, shows, collections, likes, comments, and sharing.

They also build their high-performance system as the model parameters and features server (feature store and model store). The ***Feature Store***can preserve and serve ten of millions of original features and engineered vectors. And the***Model store***will maintain and provision models and tuned\_parameters.





Online training Mechanism (Simplified Version) — Created by Catherine Wang, All Rights Reserved

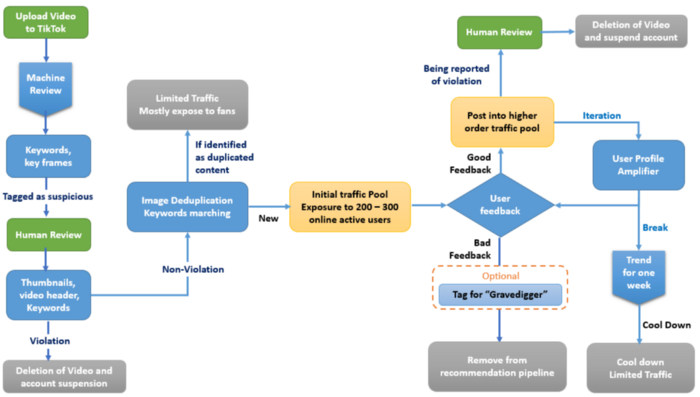
The overall training process is that 1) the online server captures the real-time data then store them into the Kafka, 2) Storm cluster consume Kafka data and product features, 3) feature store collect new features and recommendation labels to construct new training set, 4) online training pipeline retrain the model parameters, same them into the model store, 5)update the client-side recommendation list, capture new feedback (user actions) and circulate again.

3. TikTok’s Recommendation Workflow

TikTok never reveals its core algorithm to the public or the tech community. But based on the fragmented information posted via the company, and trails discovered by geeks using the reverse engineer techniques. I draw up the following conclusion.

*(Disclaimer — this is my interpretation and extrapolation, and might deviate from what TikTok is doing)*



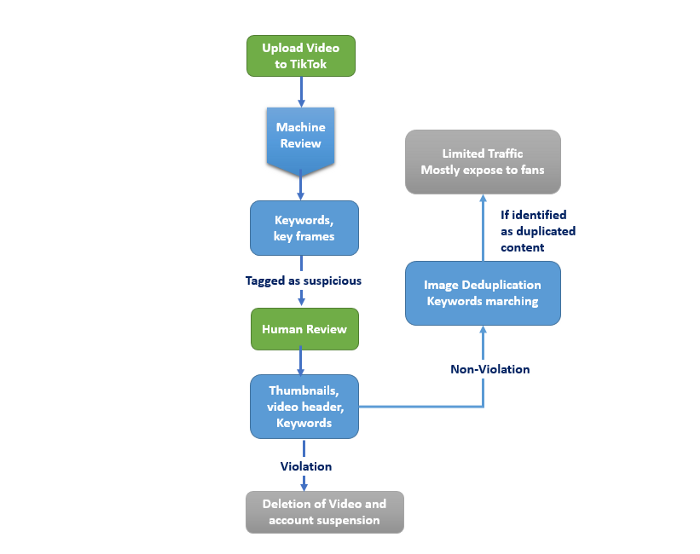


Recommendation Workflow — Created by Catherine Wang, All Rights Reserved

**Step 0: Duo-Audit system for the User Generated Content (UGC)**

At TikTok, there are millions of contents uploaded by users every day. Malicious content could easier find the loopholes in the sole machine review system, and manual review is not realistic in this context. Therefore, duo-review become TikTok primary algorithm to screen video content.





***Machine review:***Generally speaking, the Duo-audit model (computer vision-based ) can identify your video images and keywords. It mainly has two principal functions: 1) review whether there are breaches in the clips and check for the copywriting. If suspected of violation, the content will be intercepted by the model and tagged as yellow or red for human review. 2) by extracting the pictures and keyframes from the video, TikTok’s duo audit algorithm would match the extractions with its massive archived content base. Duplication will be picked up and rendered lower traffic and put less weighing the recommendation engine.

***Manual review:***mainly focused on 3 areas: Video Title, Cover Thumbnail, and Video Keyframes. For the content tagged as suspicious via the Duo-Audit model, technicians will manually review them. If identified as a violation of regulation, the video will be deleted, and suspend the account.

**Step 1: Cold Start**

The core of TikTok’s recommendation mechanism is ***Information Flow Funnel.***When content passed the duo-audit review, it will be put into a cold start traffic pool. For example, after your new video passes the review process, TikTok would assign initial traffic of 200–300 active users, there you could gain up to a few thousand exposures.

In this mechanism, a new creator can compete with a social influencer (who might have tens of thousand followers), because they have the same starting point.

**Step 2: Metric based Weighing**

Through the initial traffic pool, video can gain thousands of views, and those data will be collected and analyzed. Metrics to be considered in the analysis include likes, views, complete views, comments, followers, reposts, shares ata and etc.

Then the recommendation engine will weigh your content based on those initial metrics and your account score(whether or not you are a high-quality creator).

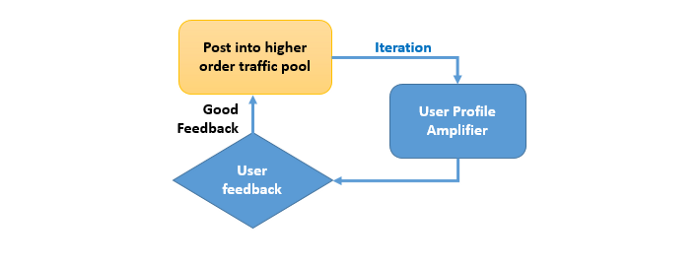
If the engine decided to weigh up your content, the top 10% will be feed in an additional 10,000–100,000 traffic exposures.

**Step 3: User Profile Amplifier**

The feedback from step 2 traffic pool will be further analyzed for the decision of using the user profile amplifier. In this step, outperforming content will be strengthened and amplified in a specific user group (e.g., sports fans, fashion lovers).

This is similar to the concept “guess what you like” function. The recommendation engine will build a user profile base so that it can find the best match between the content and the user group.





**Step 4: Boutique Trending Pool**

Less than 1% of the content will eventually enter the Trending Pool. The amount of exposure the content can get in this pool is a magnitude higher than others. Because the trending content will be recommended to all users indifferently. (Assumption, no matter who you are, you might want to see the latest protesters video of “Black lives matter”)

**Other Step: Delayed Ignition**

Some Tiktokers would notice that their content suddenly gets enormous traction after weeks of posting with average performance.

Mainly there are two reasons:

* Frist, TikTok has an algorithm (nickname “gravedigger” ) to look back to the old content and mining for the high-quality candidates for exposure. If your content has been pick by this algorithm, it’s indicating your account has enough vertical videos to derive a clean label. This label will increase your content visibility in the gravedigger.
* Second, “Trendy effect.” It means that if one of your content gets millions of views, it will direct the traffic to your main page, thus increasing the views of our old content. This often occurs in the vertical creator (e.g., funny cat video creator). One trendy video will ignite all the other high-quality videos (people want to see more of your cute, curious cat).

Limitation: Traffic Peaking

If one content cloud passes the information flow funnel (duo-audit, weighing iterations, and amplifiers), the creator’s account will gain excessive exposure, user interactions, and fans.

But this high exposure time window is narrow, based on the research. Usually, the window will last around a week. After this time slot, this content and account will cool down, and even the subsequent videos can hardly become trendy.

Why?

The main reason is TikTok wants to introduce varieties and remove unintentional bias in its algorithm. By this design, the recommendation engine won’t be inclined to a particular type of content, thus make sure that new content will get equal opportunities to get into the trendy pool.

The topic about ethics in AI algorithms and social media is more likely a ‘hot potato’ in the hands of data analyzers, software engineers and business owners. If I must be honest, nowadays the ethics is something that changes its own definition in a matter of days. Yesterday’s understandings of what is ethical are today’s obsolete news. The majority of our society lives with the myth that their personal data and space are their own only and no one must intervein with that. People think that no one has the right to collect their data, especially without their knowledge, but that is far from the truth. Every social media, doesn’t matter the flavor, collects data, and sells them to a whoever is ready to pay for that. Every successful web site collects data – some of them store user’s habits and locations, others – well they collect much more information. In the corporate world information costs a lot. The software developers, who create all these algorithms are employed by those businesses. Often, they have no other choice, but negligee their own ethical code to satisfy the requirements of the company. The product that they made reflects talents and experience, creativity, and imagination, but at the same time has been used in a questionable way from the enterprises. The enterprises don’t really care about the ethics, since they have army of lawyers that are getting paid to twist every single word from the ethical code. This applies not only to collect personal information, but also to track every kind of data flow that user can generate in the virtual space. Thousands of software developers create variety of new pieces of algorithms that can track and store everything that enters the World Wide Web. Then all this information has been analyzed and categorized by sophisticated AI clusters and used accordingly. The application of AI is not limited just to collect and analyze bytes, but also is used to create virtual profiles, or even virtual reality. Some of their abilities is to analyze media content, such as video and audio, and as result to produce fake identities and clips, spreading false information and paranoia among the internet users. This topic is so vast, that I could spend days talking about it with proofs. In conclusion, the ethics no longer exists as a moral fundament, but is solely dependent on how the rules are amended due to demands coming from the corporate world. These days everything is for sale, and this is not the computer programmers fault.

Ethical Considerations in Predictive Analytics

“MIT and Mozilla want people to better understand the disturbing power of deepfake videos in a new project called In Event of Moon Disaster.” I believe that is an old news, but still is blowing the alarm, how the society could be manipulated by AI. Don’t get me wrong – AI algorithms combined with “deep learning” models and neuron networks are not created to scare people. Their purpose is to help science and business to better understand variety of topics, that will make our progress to leap forward – like in Space technologies and travel, developing of successful models to find and sustain life on other cosmic objects, breakthroughs in medicine, including discovering of novel methods for treatments on major number diseases and so many other possibilities. That is wonderful until someone’s dark side of the mind decides to exploit these tools in unpleasant way. There is a single file line between creativity and madness and this case is called interpretation. That is very similar to the difference between a medicine and a poison – the dose! I would like to believe that all new innovations are meant only for the progress of the humanity, but this must be defined through the moral and ethical prism, so the interpretation of the achieved data is not twisted or manipulated. The MIT experiment is not the first of its kind, because 20 years ago I witnessed a similar event in my country, when a group of students created a 10-minute fake video using video algorithms and statistical methods from a very old publication, named “The theory of games”. That video was emitted nationwide and simply informed the audience about a nuclear incident in our only nuclear power plant. The effect of that simulation was devastating, causing panic among the nation, even the official authorities were convinced that something bad happened. It took them hours until they officially announce on the national TV that all was just a fake news. If I compare that old case with the MIT experiment, I could easily see the differences in the technology used 20 years ago and nowadays – way more sophisticated, but even primitive put the whole nation into emergency mode. That is a simple example what modern technologies are able to achieve when the ethics is thrown out the window.