GDC × NetEase Games Passion of gamers

March 21-25,2022
San Francisco,CA

NeteaseGames AI Lab

Aegis Engine: Building Multi-modal Moderation System in NetEase Games

Yunbo Peng
Senior AI Engineer
NetEase Games AI Lab

#GDC22
#NetEase Games

About

- Established in 2017
- Apply AI technology to games
- Research Interests: CV, NLP, RL and Speech Signal Processing

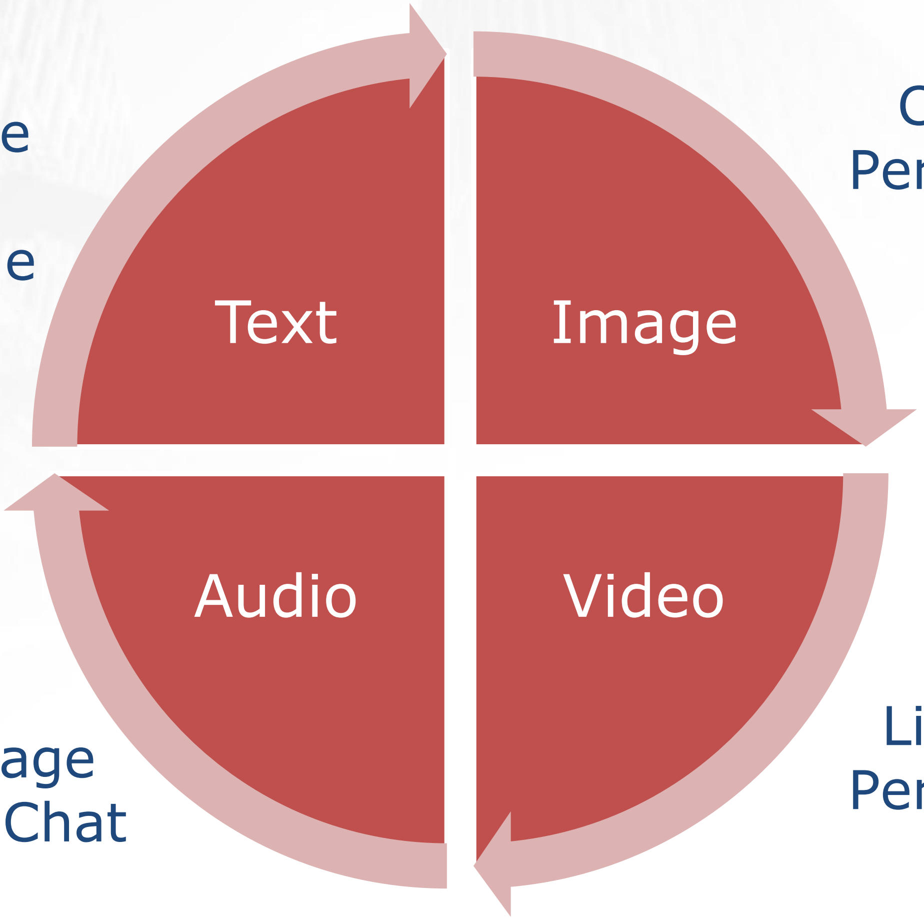
NetEase Games

Netease Games AI Lab

Outline

- **Introduction**
- Image Sub-system
- Audio Sub-system
- Text Sub-system
- Application

Introduction

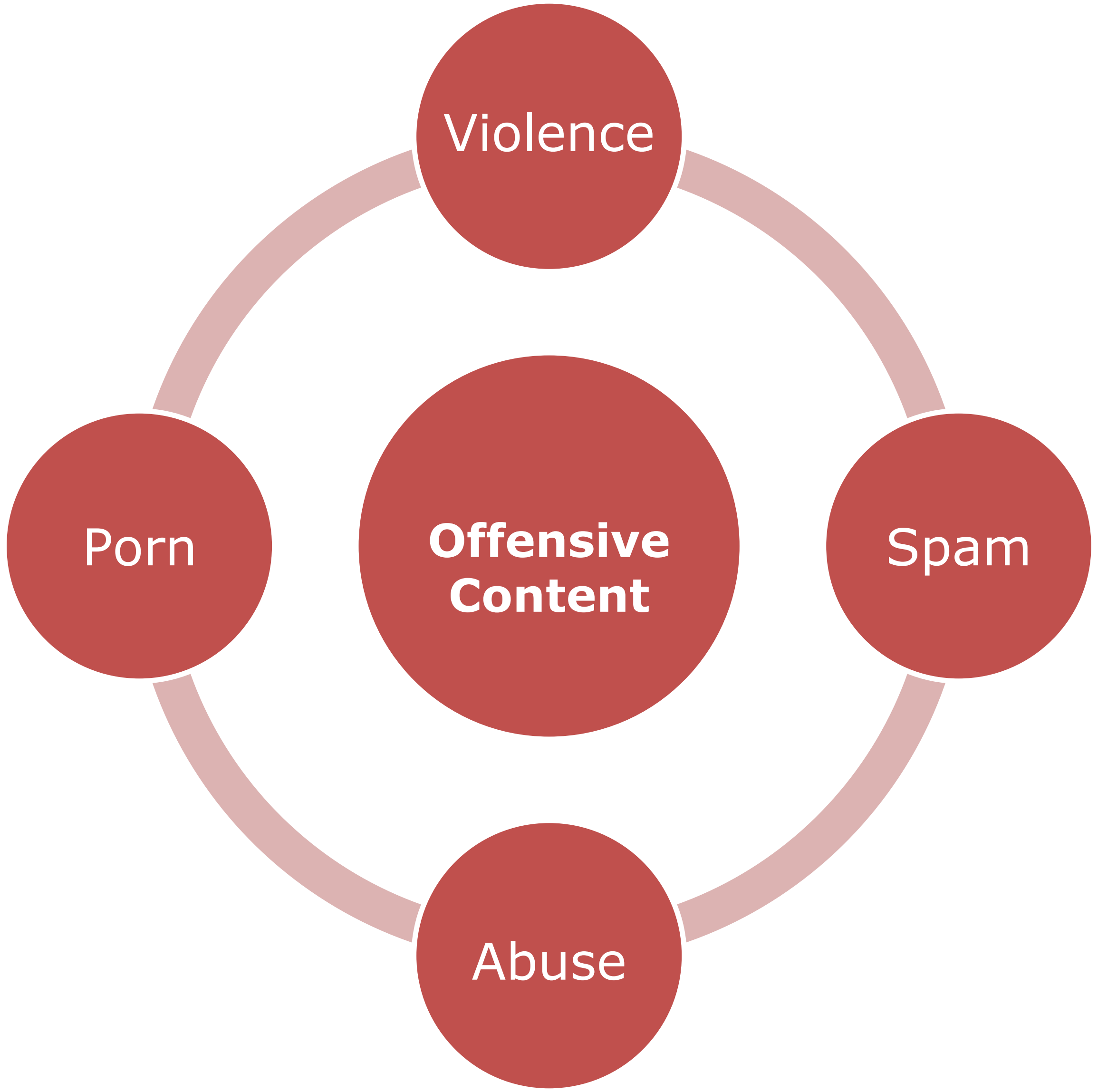


Chat Message
Nickname
Status Tagline

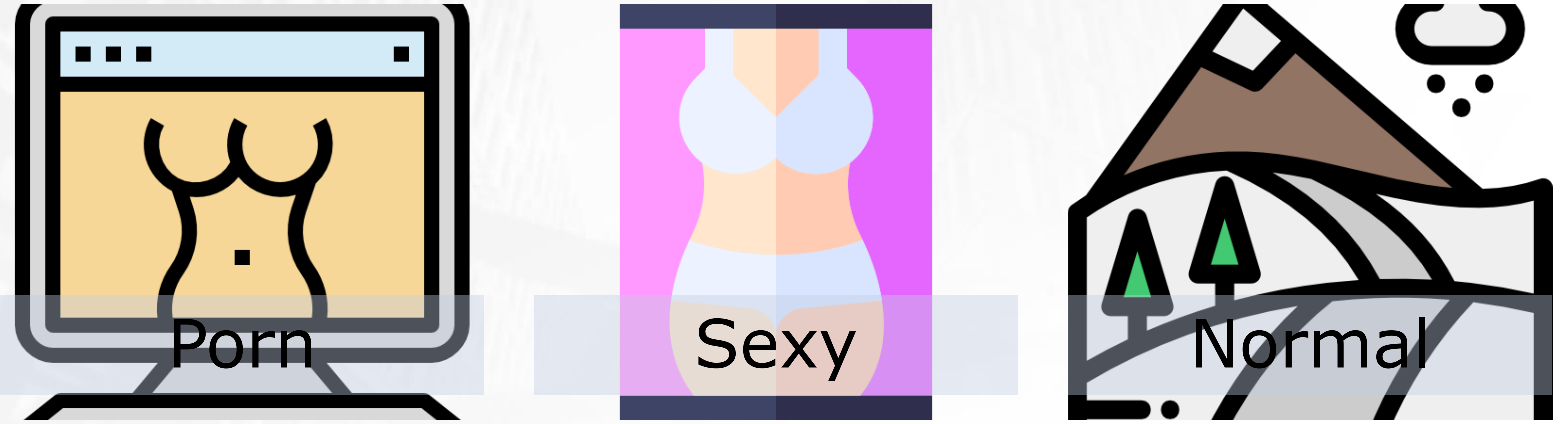
Chat Message
Personal Timeline
Forum Post

Voice Message
Team Voice Chat

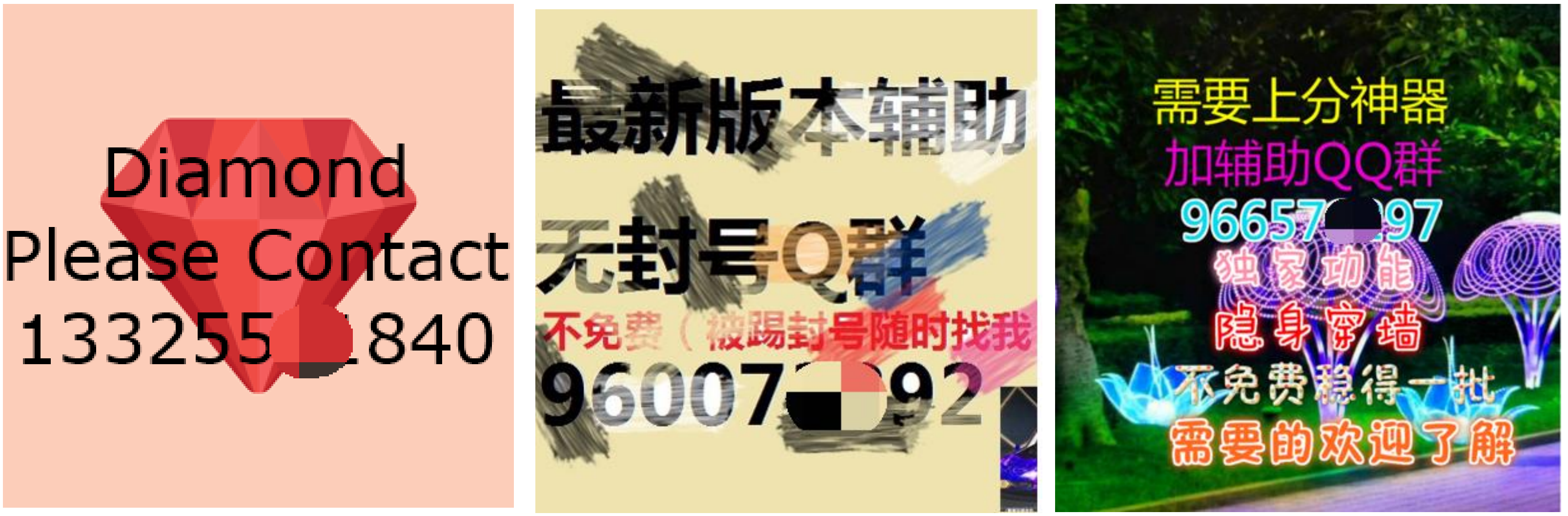
Live Streaming
Personal Timeline
Forum Post



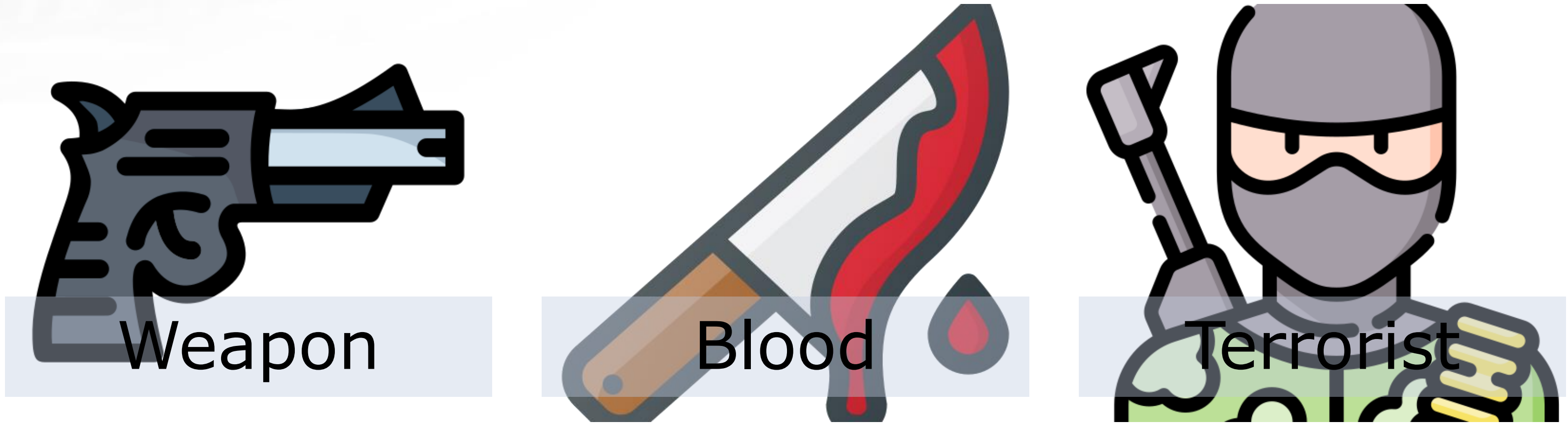
Introduction



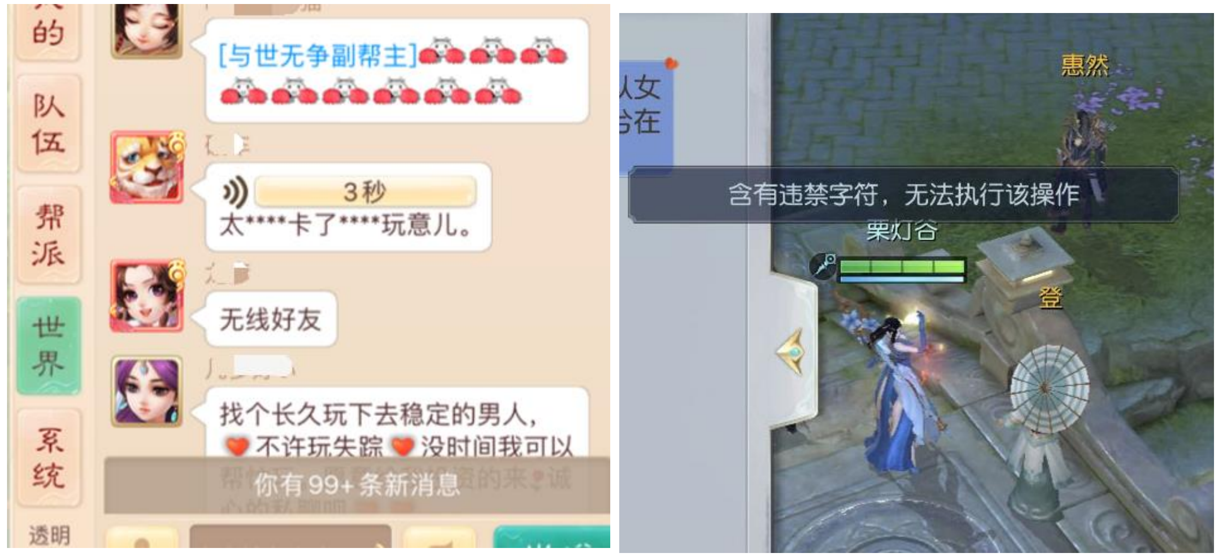
Porn



Spam



Violence



[与世无争副帮主]

3秒
太****卡了****玩意儿。

无线好友

找个长久玩下去稳定的男人，
❤不许玩失踪❤没时间我可以

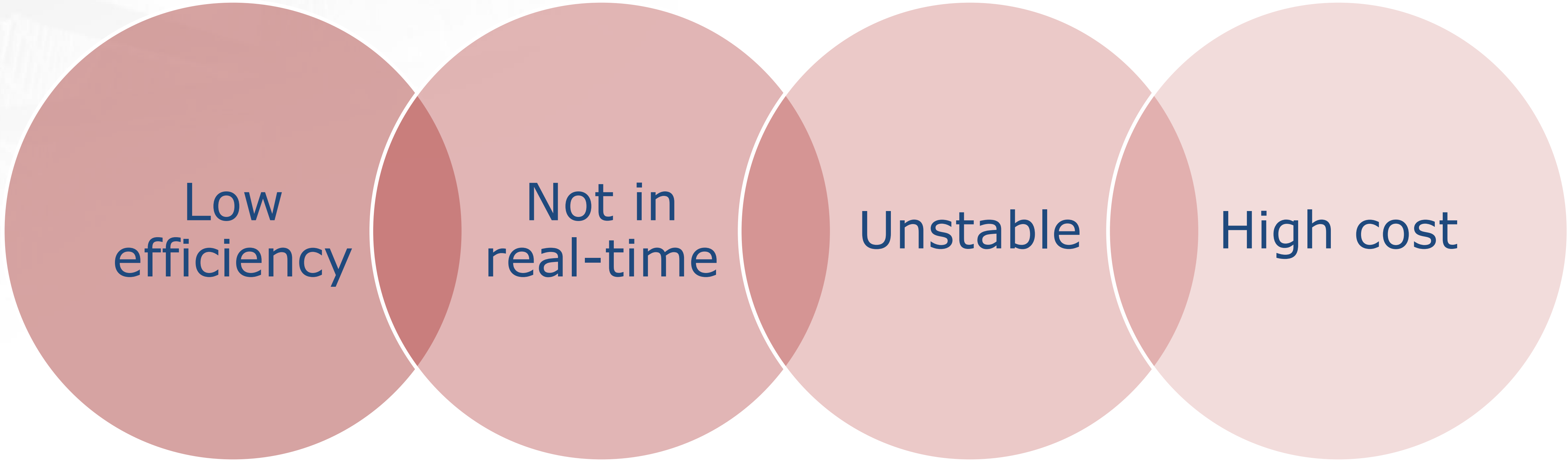
你有99+条新消息

含有违禁字符，无法执行该操作

Abuse

Introduction

Manual moderation



Low efficiency

Not in real-time

Unstable

High cost

Slow manual processing

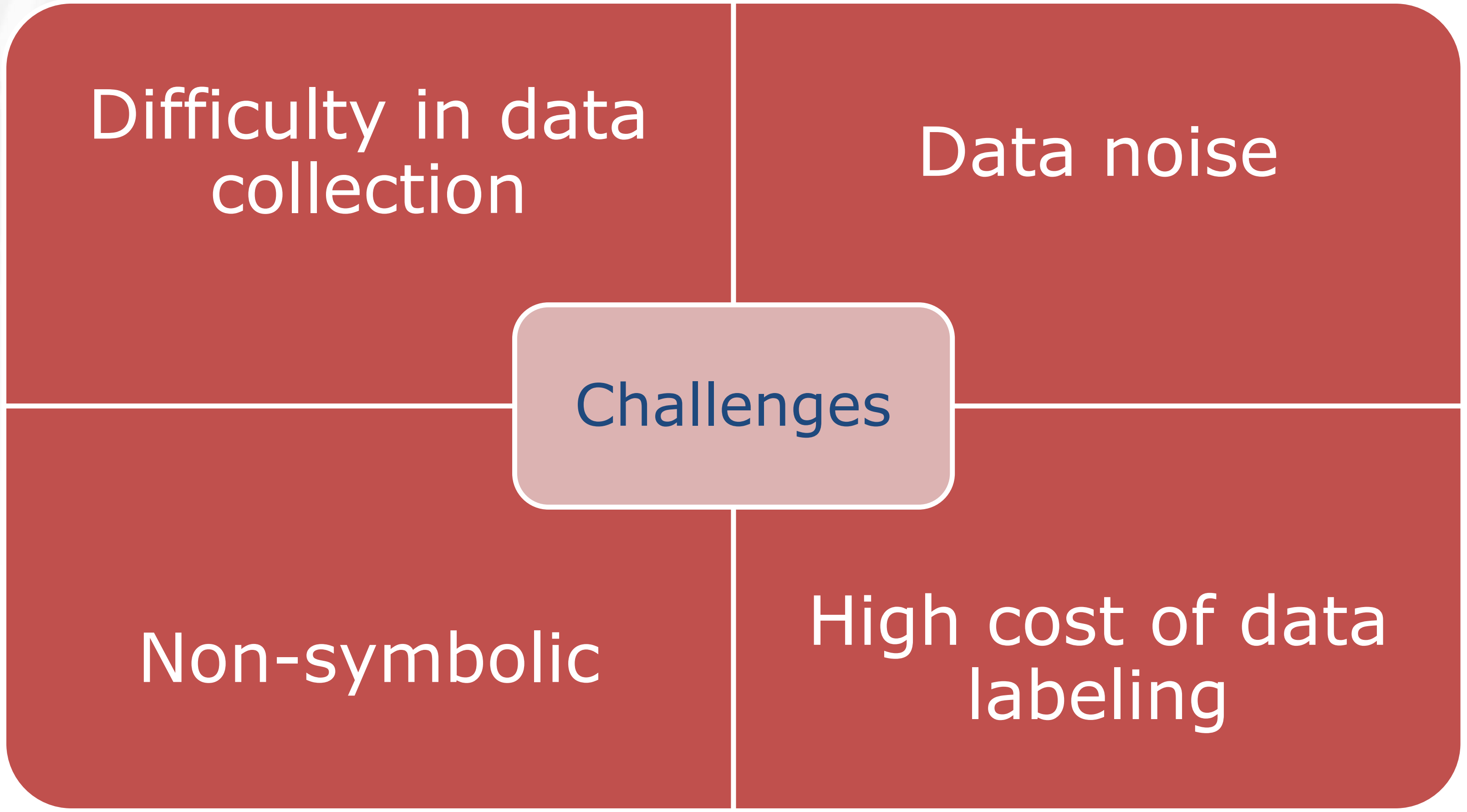
No 24/7 or immediate response

Inconsistent standards

Requires many moderators

Introduction

Challenges during development



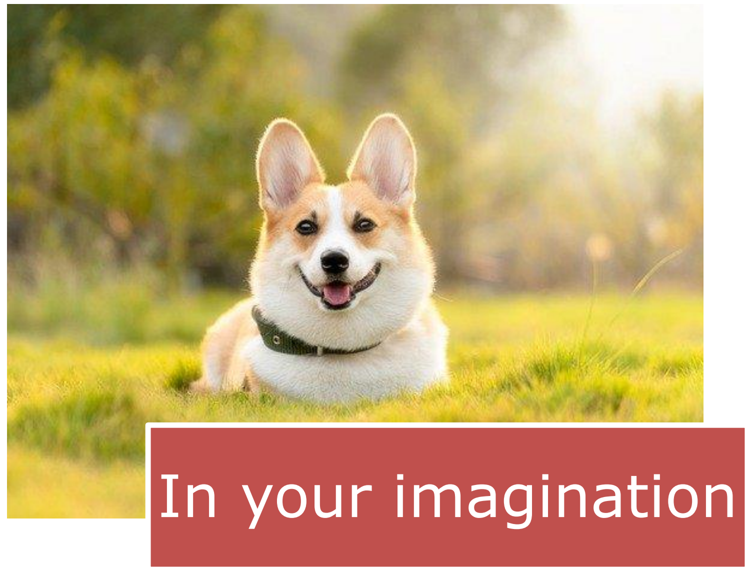
Difficulty in data collection

Data noise

Challenges

Non-symbolic

High cost of data labeling



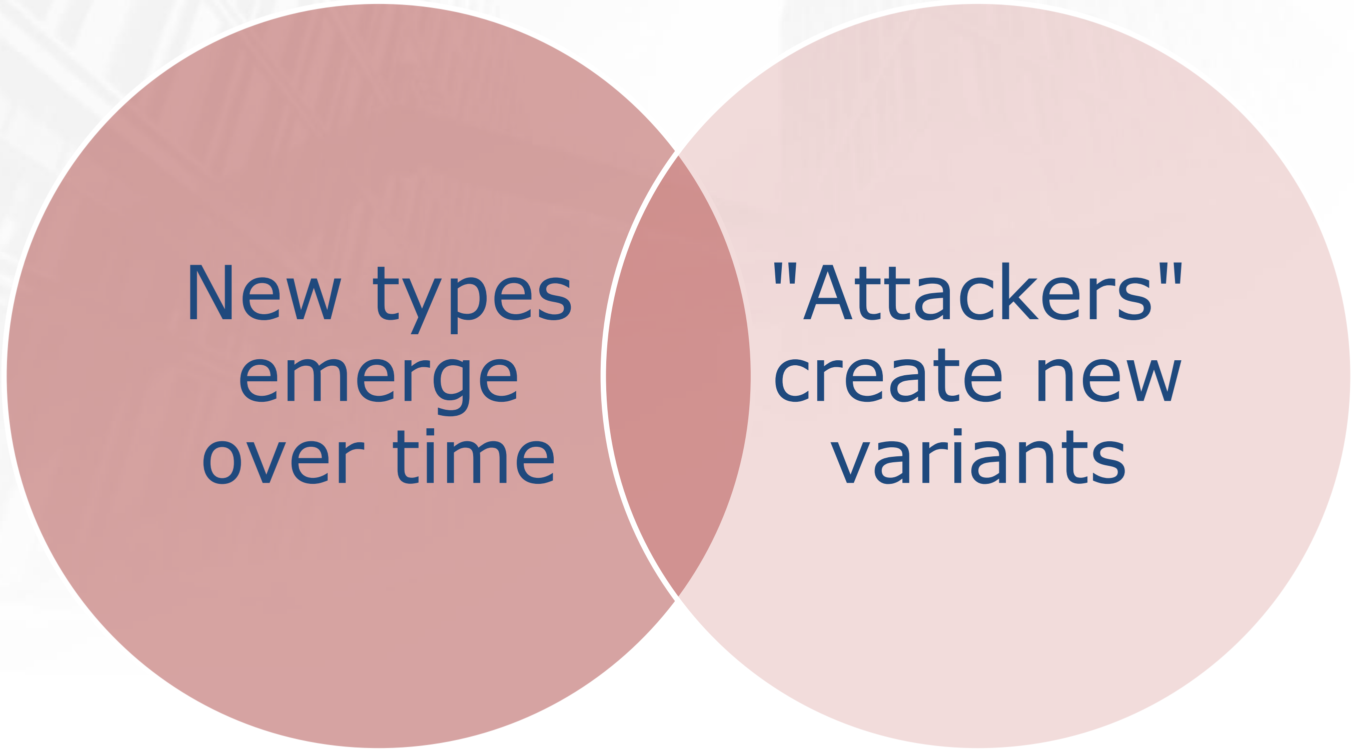
In your imagination

The actual situation

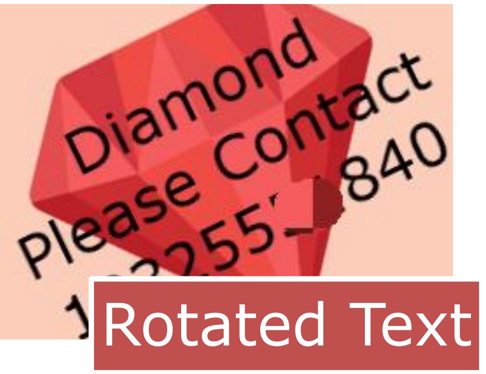
Non-symbolic image

Introduction

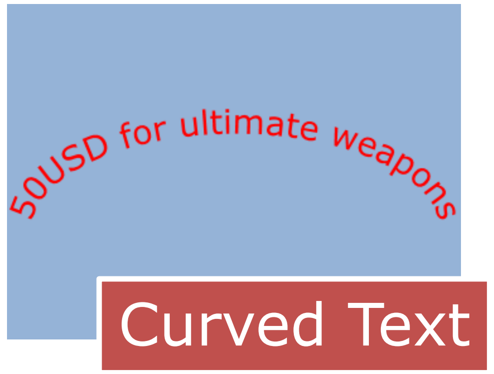
Challenges during maintenance


New types emerge over time

"Attackers" create new variants



Rotated Text



Curved Text

Handwriting

New variants of image spam

Outline

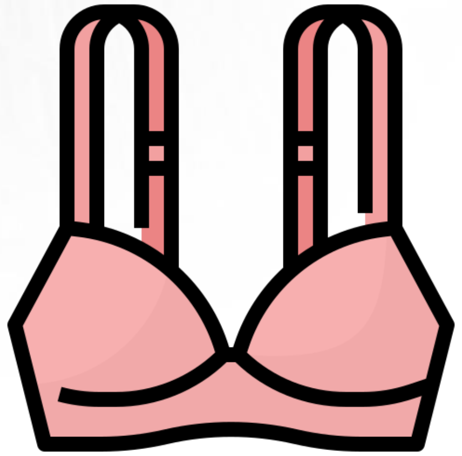
- Introduction
- **Image Sub-system**
- Audio Sub-system
- Text Sub-system
- Application

Define Categories

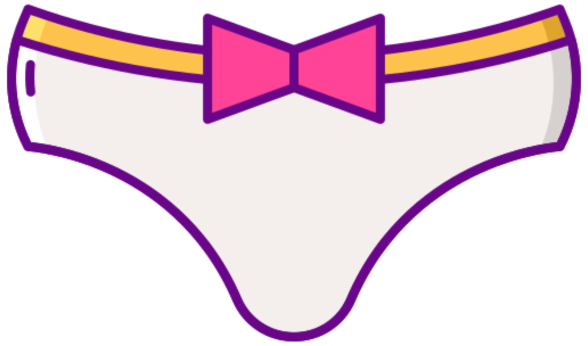
- Create categories for common scenarios and ignore outliers
- Define the scope of each category for precise annotation
- Use fine-grained categories to cover more online cases

Define Categories

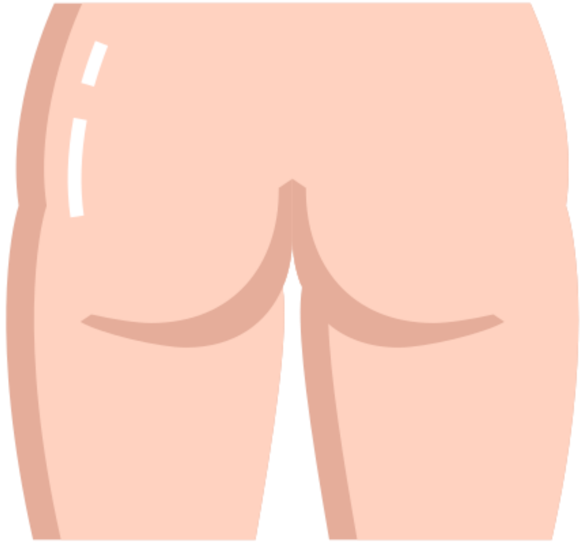
Cleavage



Bra



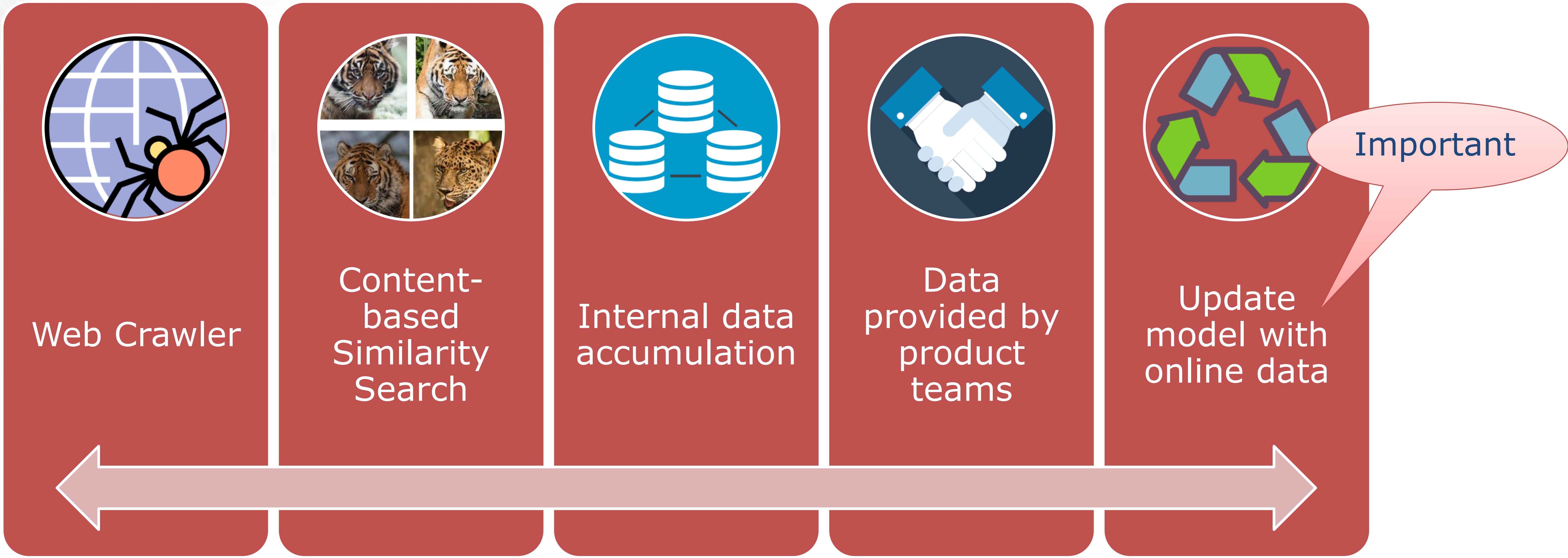
Underwear



Hip

Fine-grained categories for live streaming

Data Collection



Web Crawler

Content-based Similarity Search

Internal data accumulation

Data provided by product teams

Update model with online data

Important

Data Cleaning

Duplicated Data Removal

- MD5
- pHash/aHash…

Simple Filtering

- Traditional Features
- Data Labeling & SVM/LR Classifiers

Data Clustering

- FC Features Clustering
- Select dense and compact classes

Crowdsource Labeling

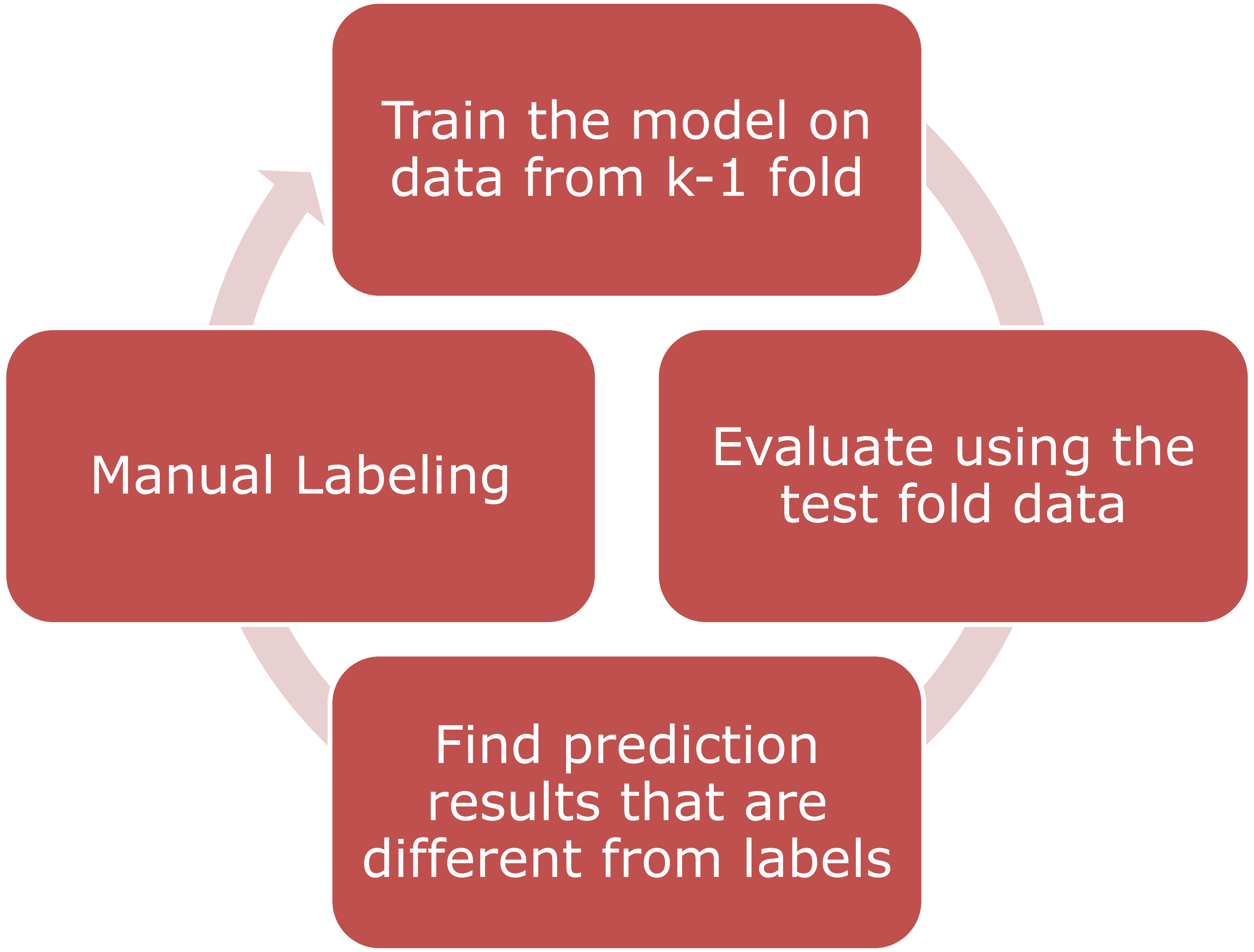
- Vote by different tools/ models
- Manual labeling

Deep Cleaning

- K-fold Cross Validation
- Inconsistent Predictions Labeling

Data Cleaning

Test	Train	Train	Train	Train
Train	Test	Train	Train	Train
Train	Train	Test	Train	Train
Train	Train	Train	Test	Train
Train	Train	Train	Train	Test



Train the model on data from k-1 fold

Evaluate using the test fold data

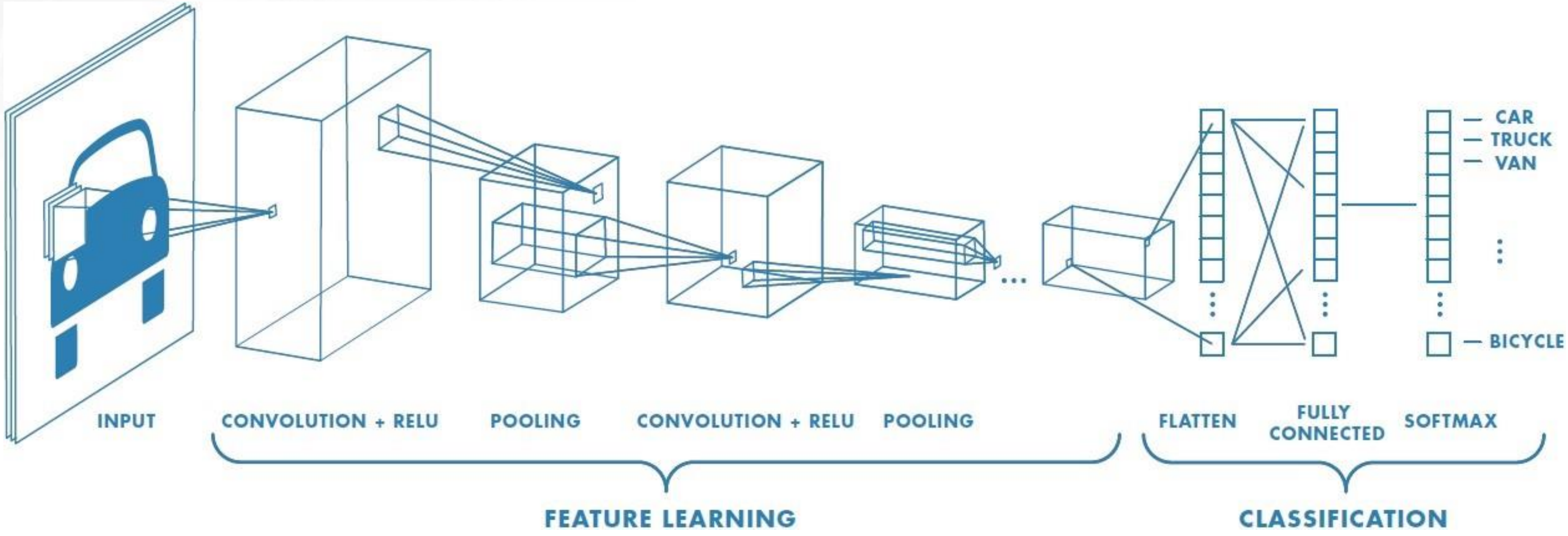
Find prediction results that are different from labels

Manual Labeling

5-fold cross validation to get deep cleaned data

Methods

- Image Classification Networks: ResNet、MobileNet v2/v3



[1] Image source: https://penseeartificielle.fr/mobilenet-reconnaissance-images-temps-reel-embarque/

Methods

Data Augmentation

- Auto Augment
- Cutout

Model Overfitting

- Dropout after GAP Layer
- Label Smoothing

Imbalanced Classes

- Over-sampling
- Weighted Loss

Model Optimization

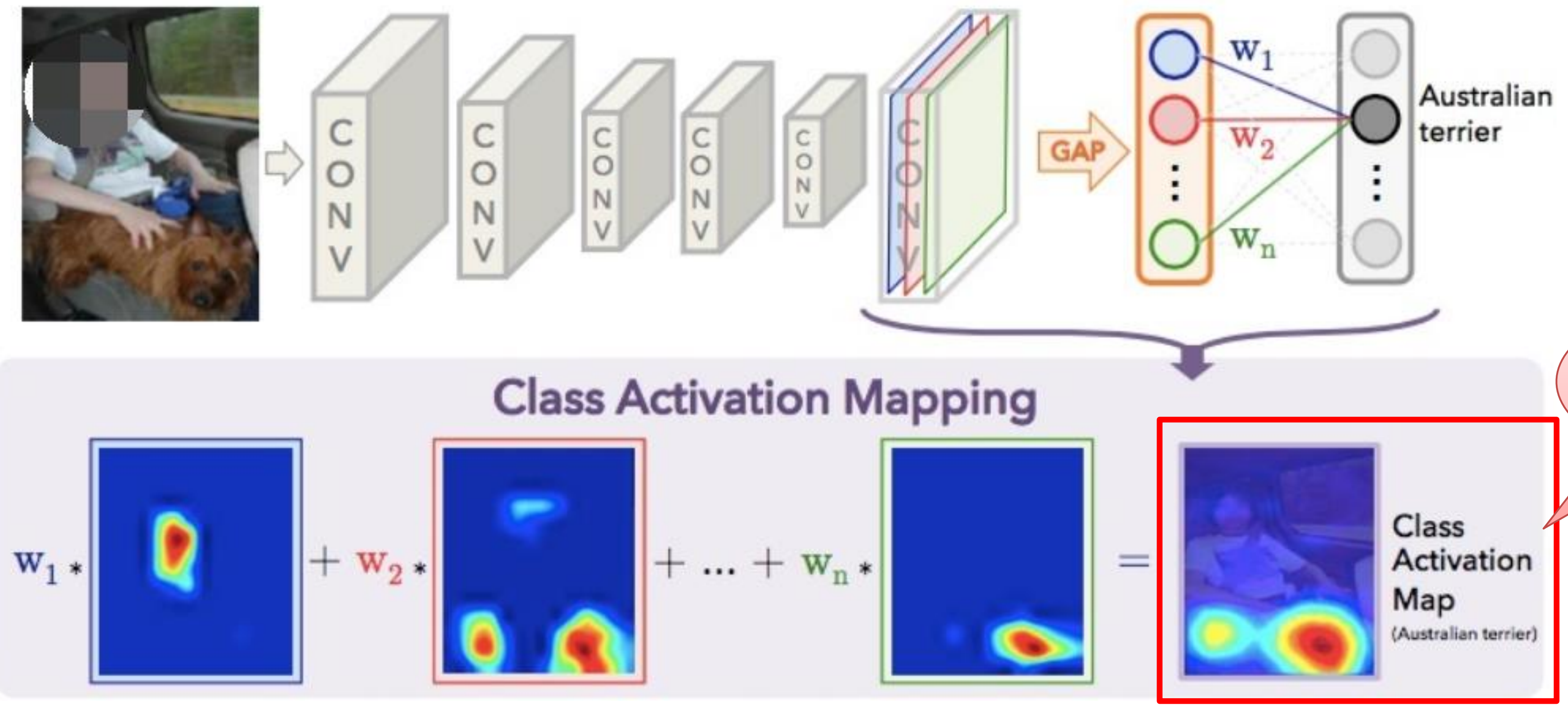
Class Activation Map (CAM)

Big Input, Small Network

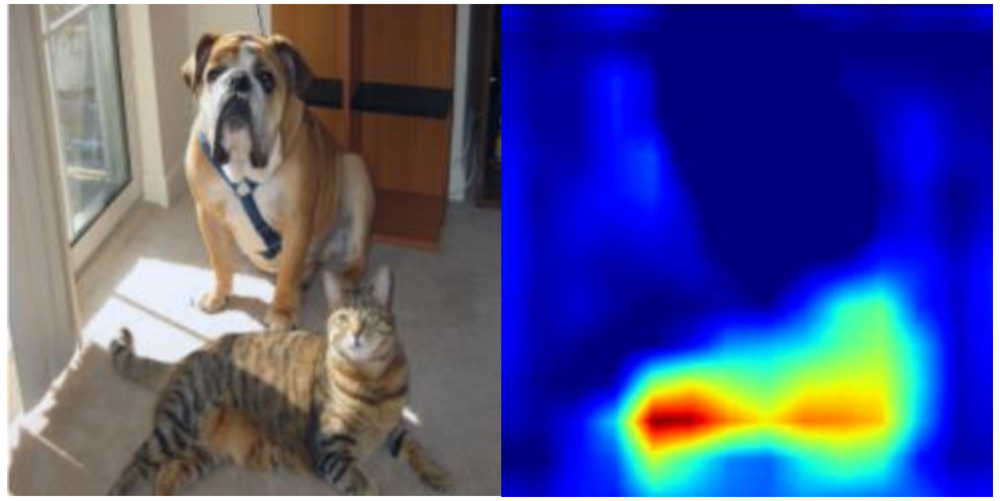
Attention Mechanism

Object Detection Networks with FPN: YOLO、RetinaNet

Model Optimization



Result

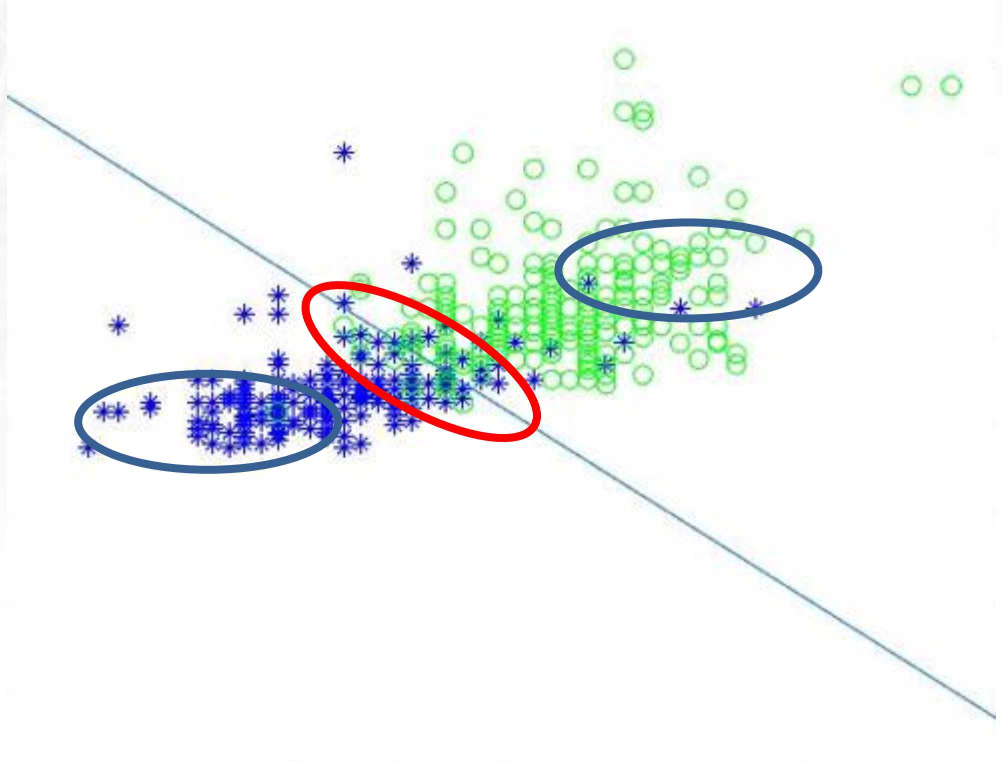


Why the prediction result is **not a dog**?
Ah, the reason is **the influence of cat**.
So we need **collecting more images containing cat**

Use Class Activation Map(CAM) to Improve Data

[1] Zhou B, Khosla A, Lapedriza A, et al. Learning deep features for discriminative localization.

Model Update



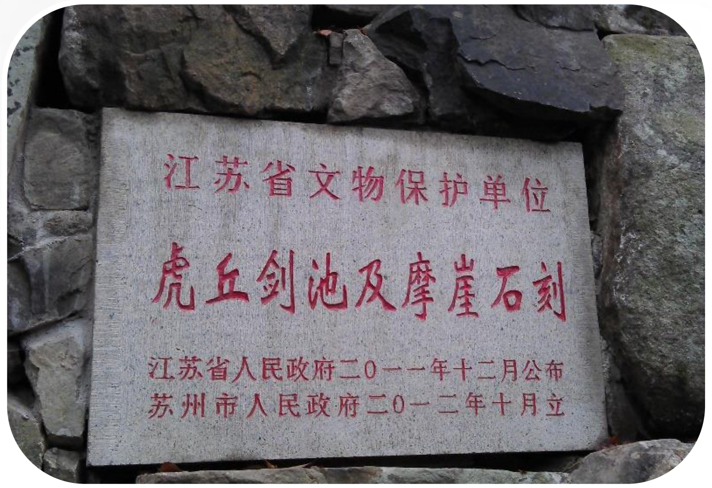
⬭ Non-ambiguous data:
Used to increase data diversity
and model stability

⬭ Boundary data:
Used to update the data boundary
and increase model robustness

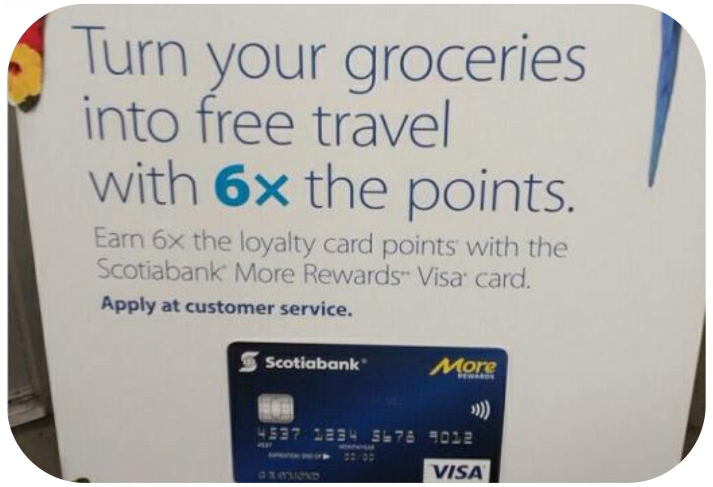
Pulling Additional Online Data to Update the Models

Scene Text Detection

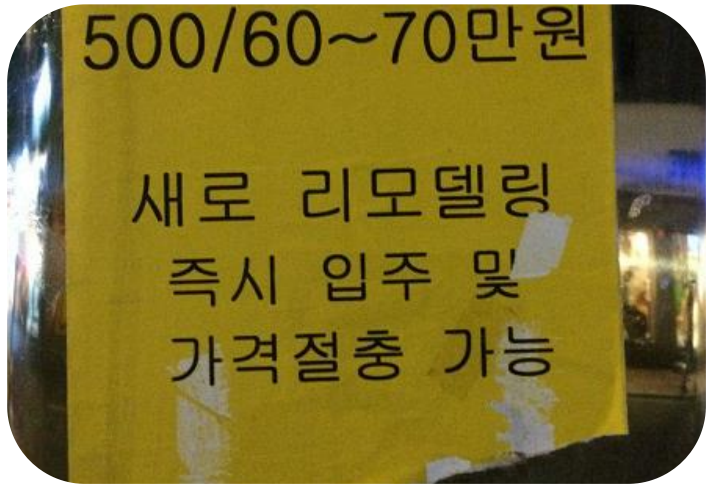
- Self-labeling datasets and public datasets
- A dataset containing 63,000+ images was constructed


江苏省文物保护单位
虎丘剑池及摩崖石刻
江苏省人民政府二0一一年十二月公布
苏州市人民政府二0一二年十月立

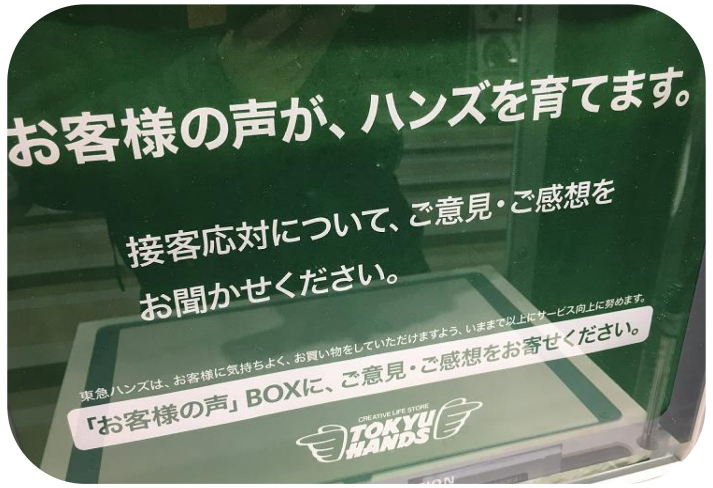
Chinese


Turn your groceries into free travel with 6x the points.
Earn 6x the loyalty card points with the Scotiabank More Rewards Visa card.
Apply at customer service.
Scotiabank
More
VISA

English



Korean


お客様の声が、ハンズを育てます。
接客応対について、ご意見・ご感想を
お聞かせください。
「お客様の声」BOXに、ご意見・ご感想をお寄せください。
TOKYU HANDS

Japanese

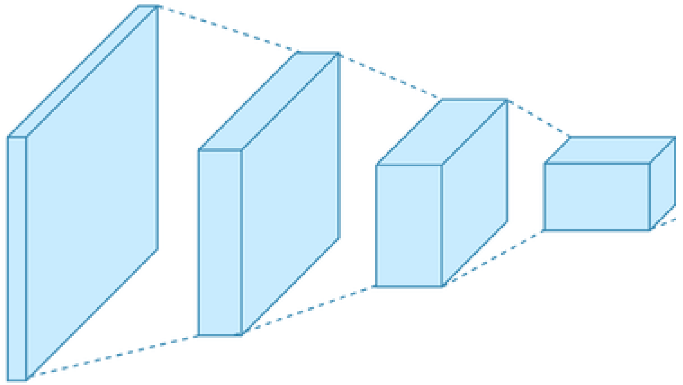
Scene Text Detection

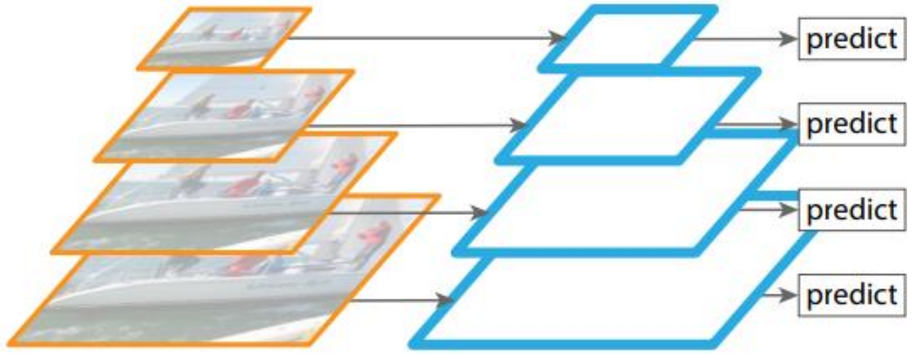
Feature Extraction Network → Semantic Segmentation → Post Processing

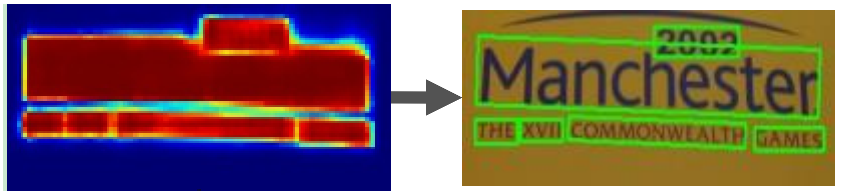
MobileNet v3、ResNet…

FPN、UNet…

Find Coutours、
Min Area Rectangle…

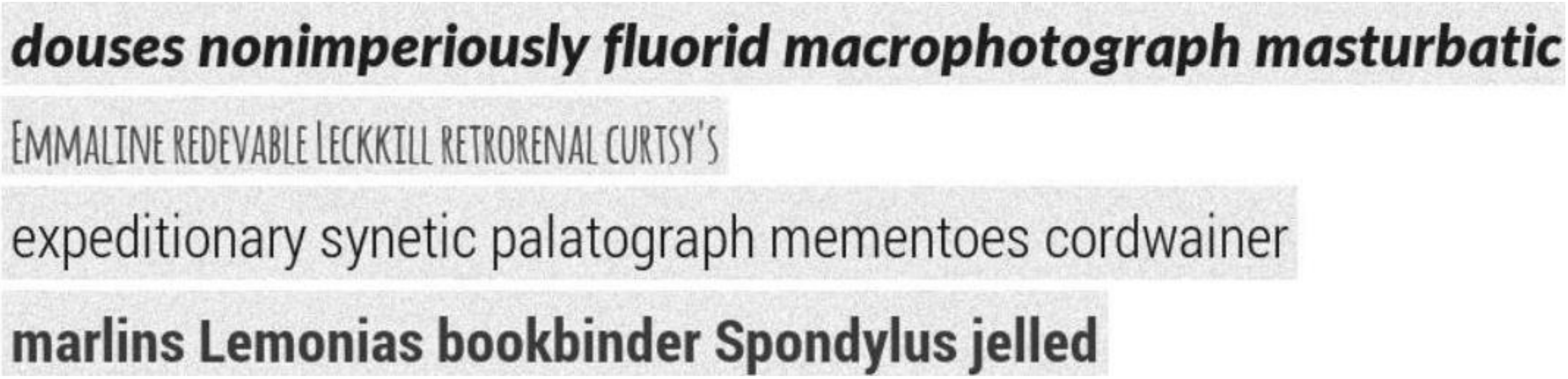






Scene Text Recognition

- Synthesis dataset
- Real-world dataset annotated by crowdsource labeling of several tools/models



douses nonimperiously fluorid macrophotograph masturbatic

EMMALINE REDEVABLE LECKKILL RETRORENAL CURTSY'S

expeditionary synetic palatograph mementoes cordwainer

marlins Lemonias bookbinder Spondylus jelled

Scene Text Recognition

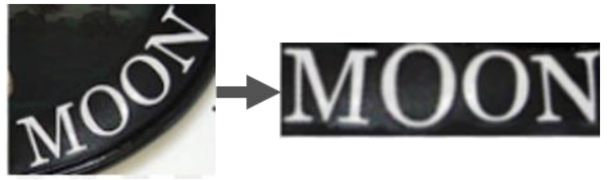
Pre Processing Network → Image Feature Extraction Network → Sequential Feature Extraction Network → Decoder

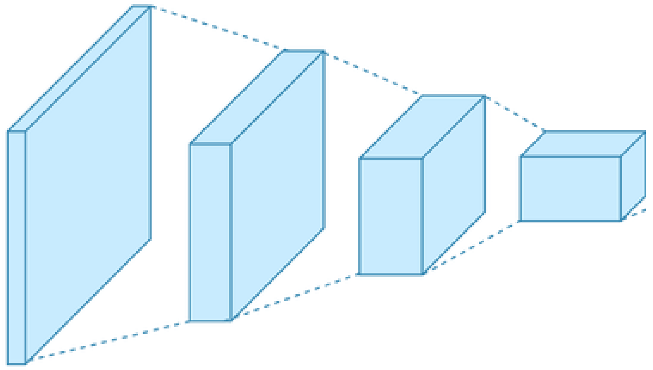
STN、TPS…

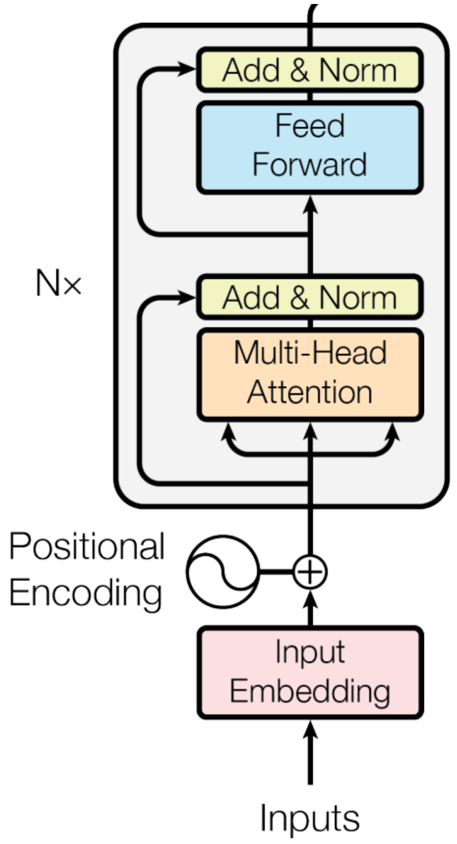
ResNet、RepVGG…

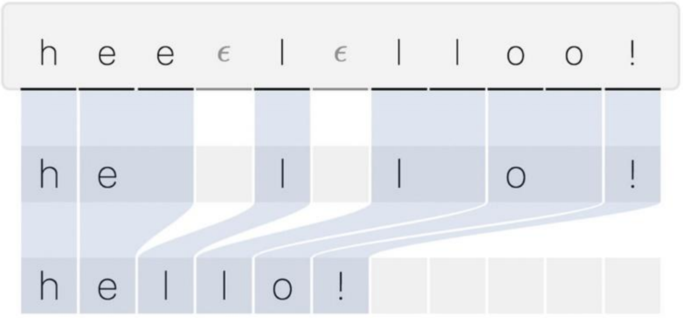
Transformer Encoder、LSTM…

CTC、Attention…









Outline

- Introduction
- Image Sub-system
- **Audio Sub-system**
- Text Sub-system
- Application

Speech Recognition

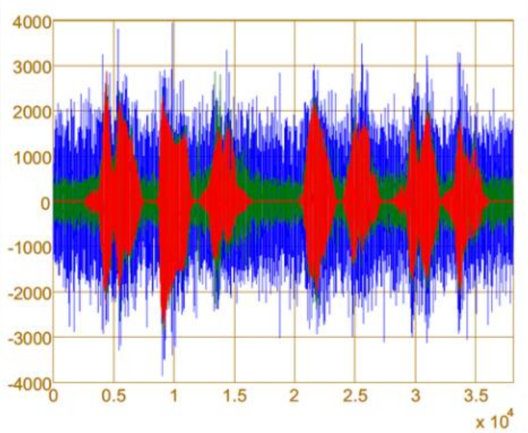
Feature Extraction Module → Acoustic Neural Network → Language Model → Post Processing

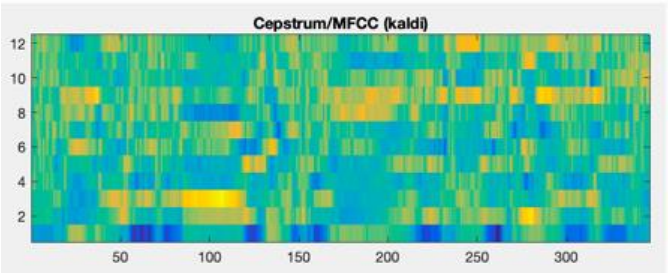
Denoise&VAD、MFCC ...

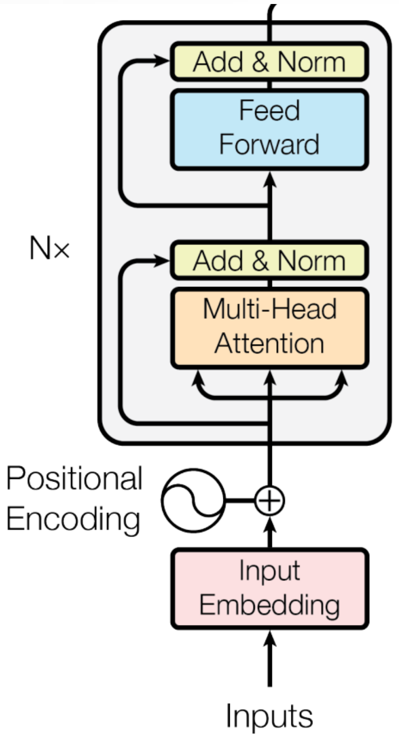
Conformer、TDNN-LSTM...

N-gram、Neural Network...

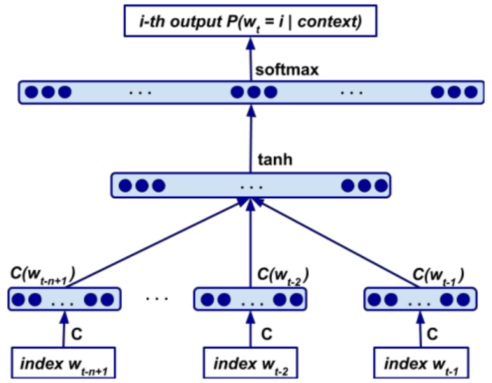
Text Filtering
Text Classification...

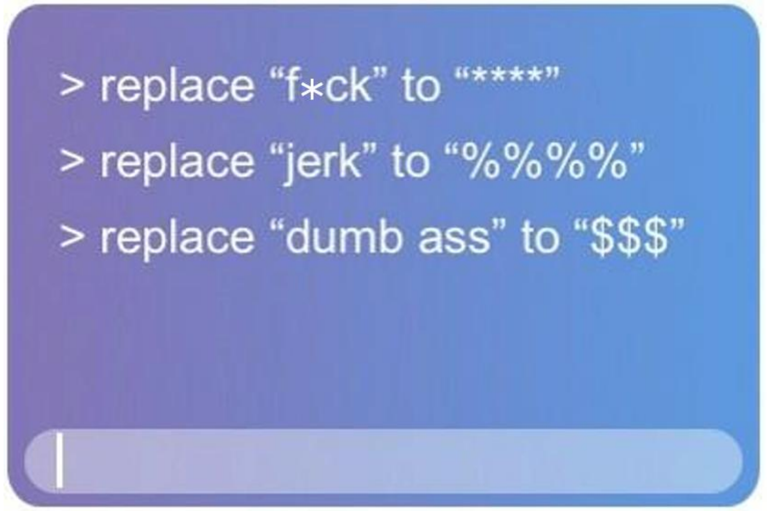






```
2-grams:
-0.7782 as the          -0.2717
-0.4771 at all          0.0000
-0.7782 at the          -0.2915
3-grams:
-2.4450 in the lowest
-0.5211 in the middle
-2.4450 in the on
```





> replace “f∗ck” to “****”

> replace “jerk” to “%%%%”

> replace “dumb ass” to “$$$”

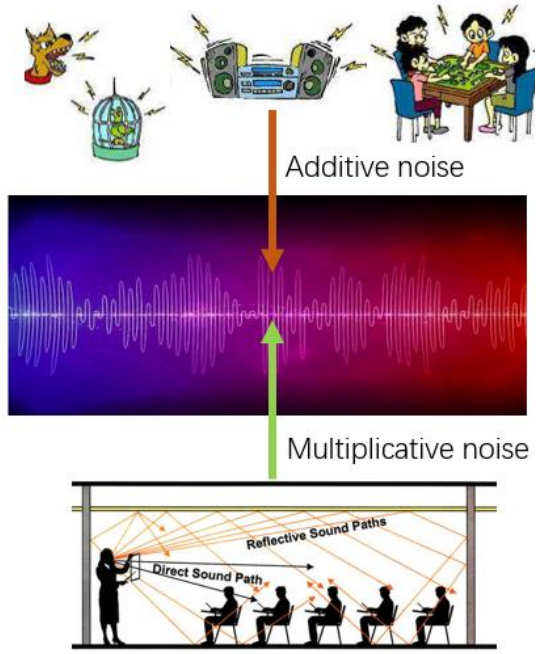
Challenge

THE SINITIC LANGUAGES

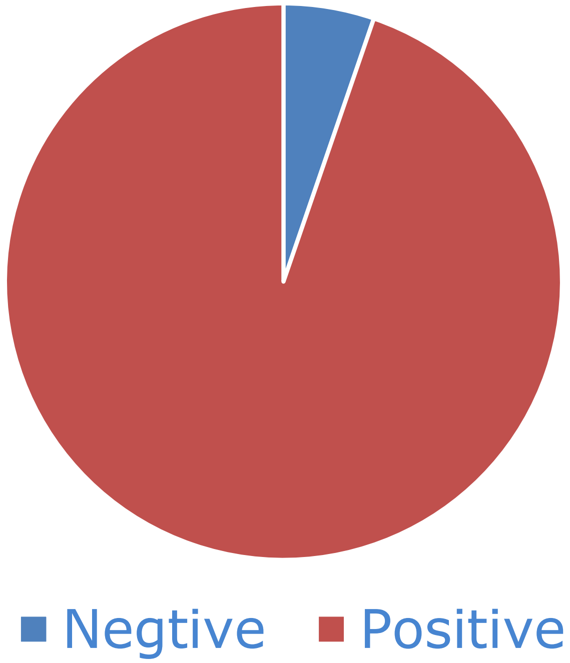
Mandarin	836 million (worldwide)
Jin	45 million(usu. grouped with Mandarin)
Wu	77 million
Hui	3.2 million(usu. grouped with Wu)
Gan	31 million
Xiang	36 million
Min	60 million(incl. Taiwanese)
Hakka	34 million (worldwide)
Yue	71 million (worldwide)
Ping	2 million(usu. grouped with Yue)

Diverse Accents

Multiple Devices & OS

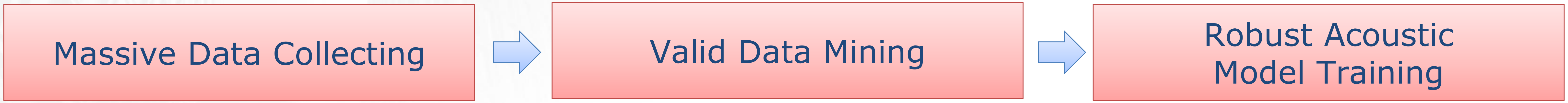


Noisy Environment



Negative Samples Sparsity

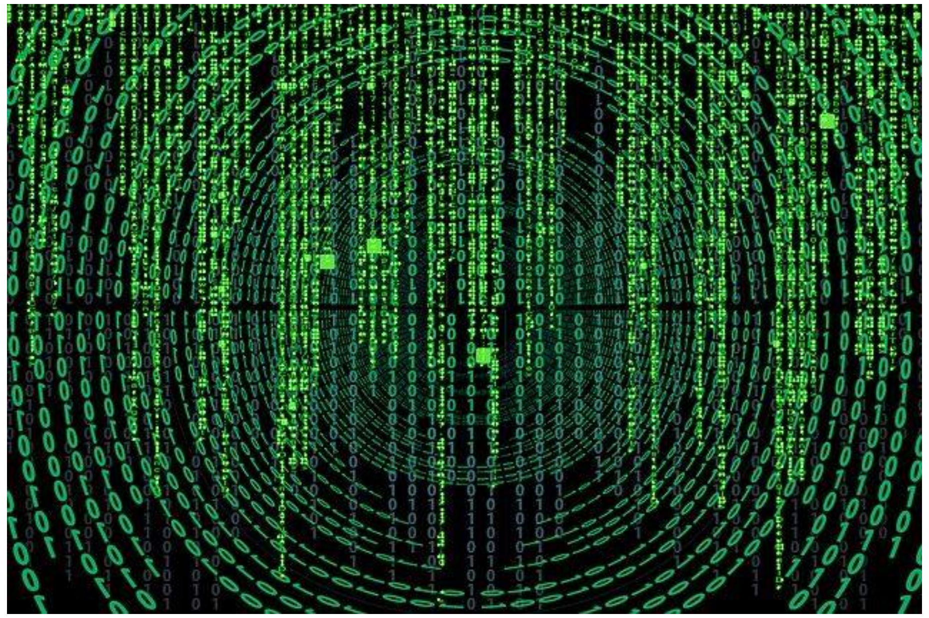
Solution



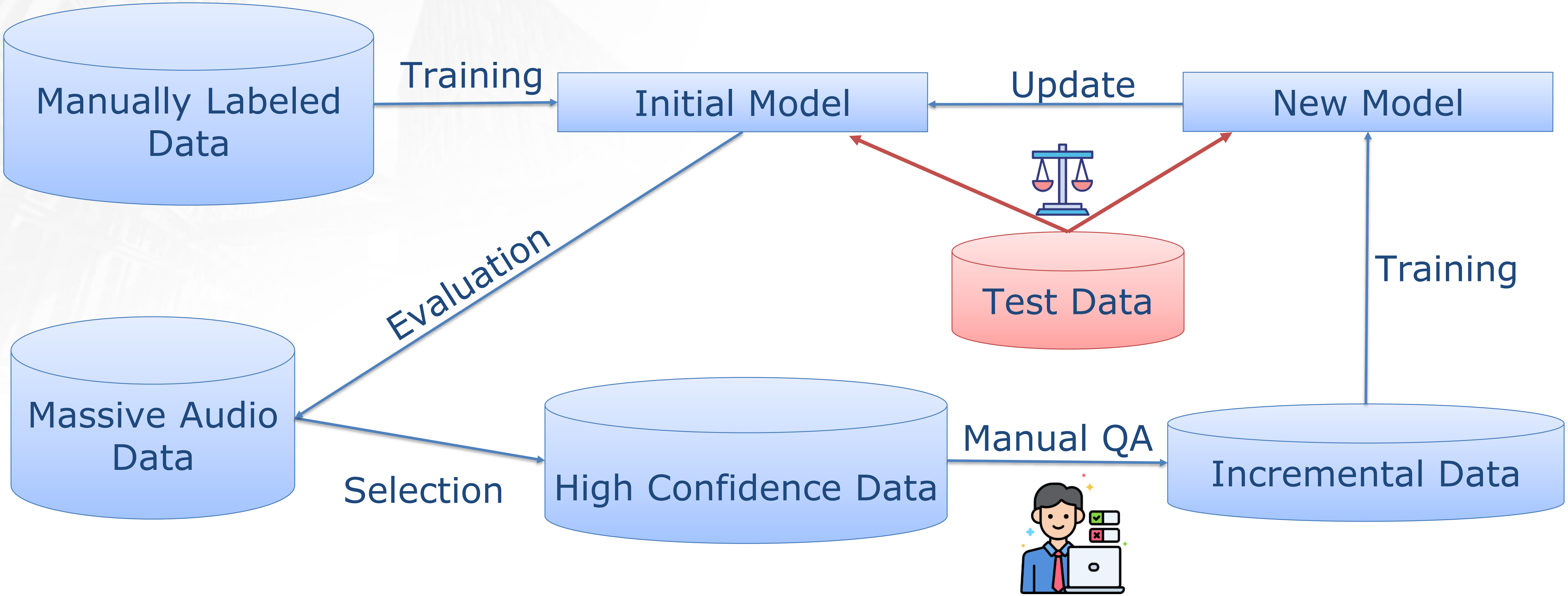
Massive Data Collecting → Valid Data Mining → Robust Acoustic Model Training

100,000+ hours
accents/devices coverage…

Closed-loop
Incremental…

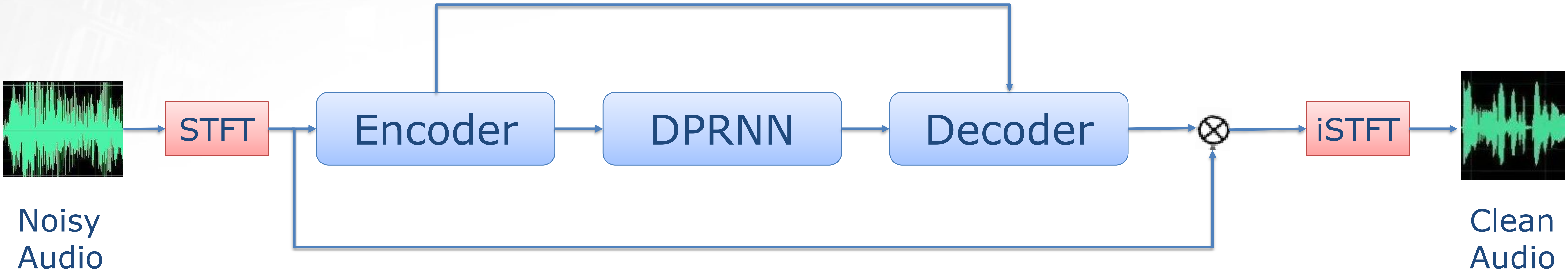


Valid Data Mining



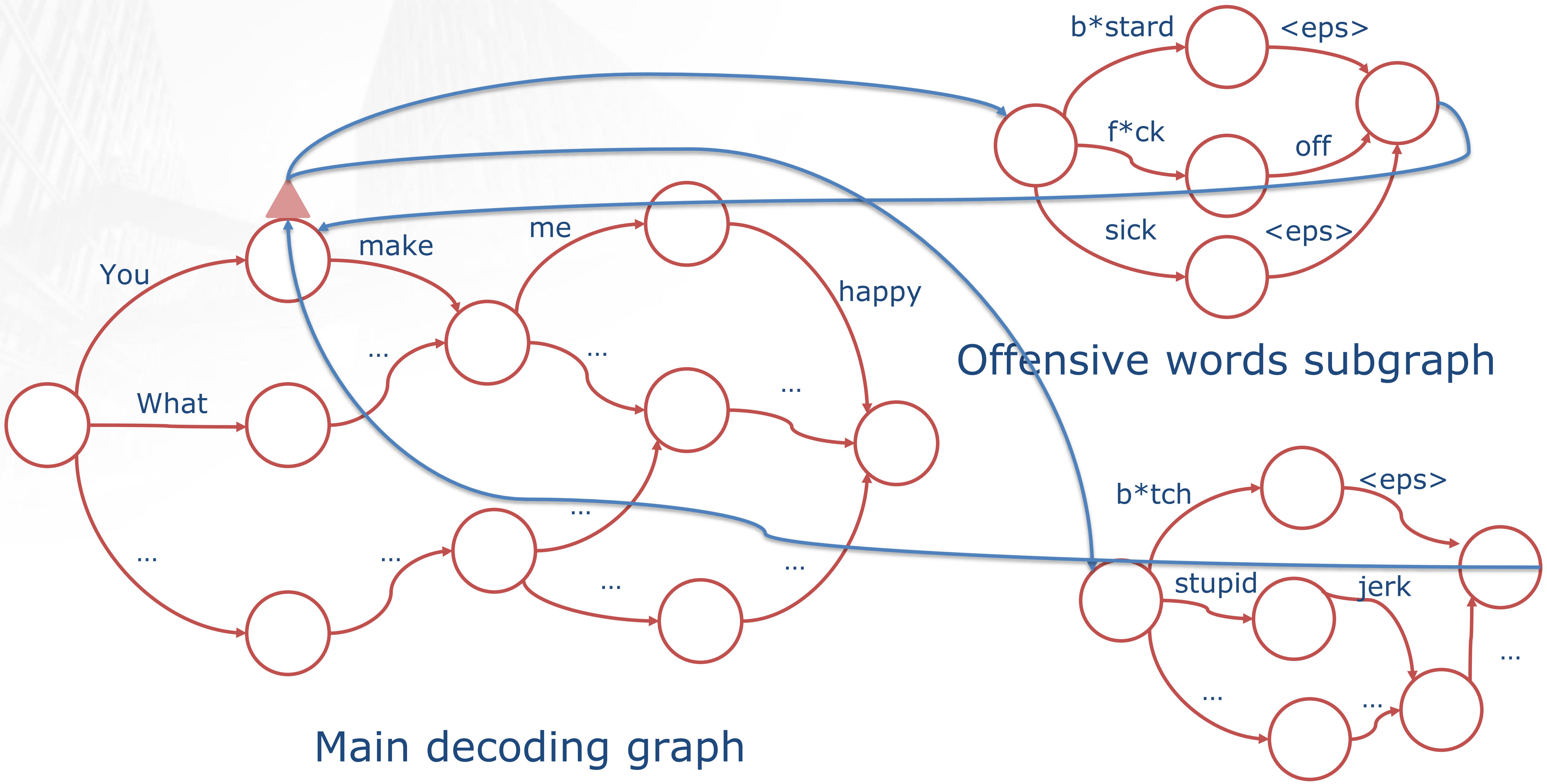
Noise Reduction

- Simulation data in 3000+ scenes
- Following DPCRN Network[1] but more lightweight



[1] Le X, Chen H, Chen K, et al. DPCRN: Dual-Path Convolution Recurrent Network for Single Channel Speech Enhancement.

Keyword Enhancement



Offensive words subgraph

Main decoding graph

New offensive words subgraph

Outline

- Introduction
- Image Sub-system
- Audio Sub-system
- **Text Sub-system**
- Application

Text Filtering

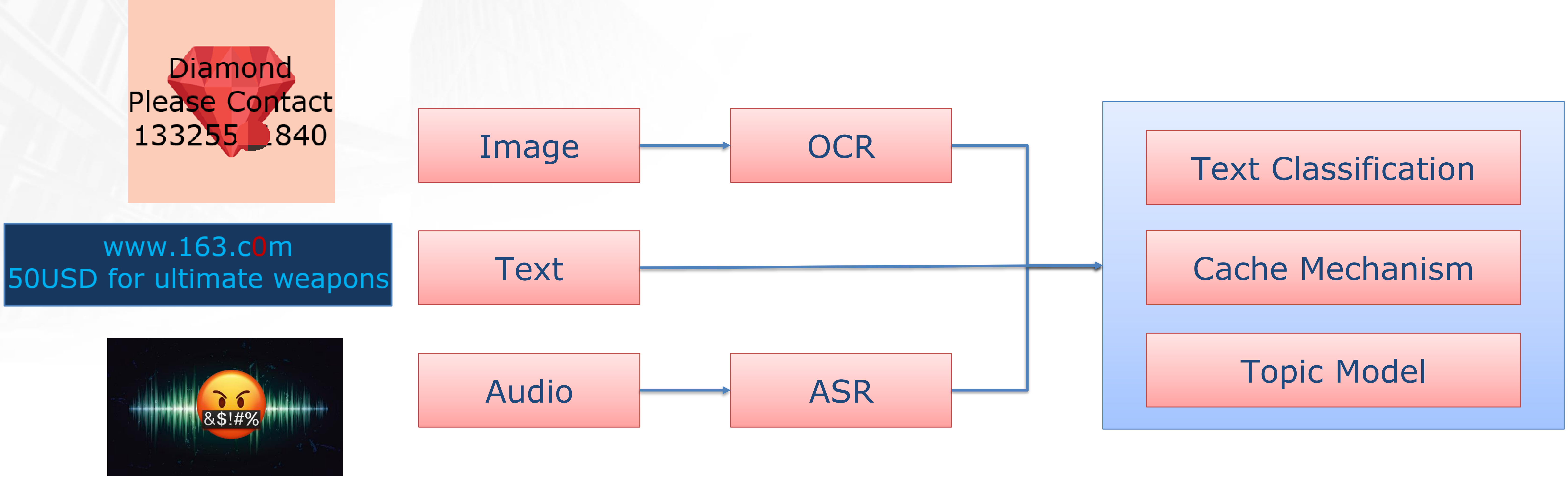
www.163.c0m
USD|RMB|EUR

Regular Expression

MACHINE LEARNING

Machine Learning

Architecture

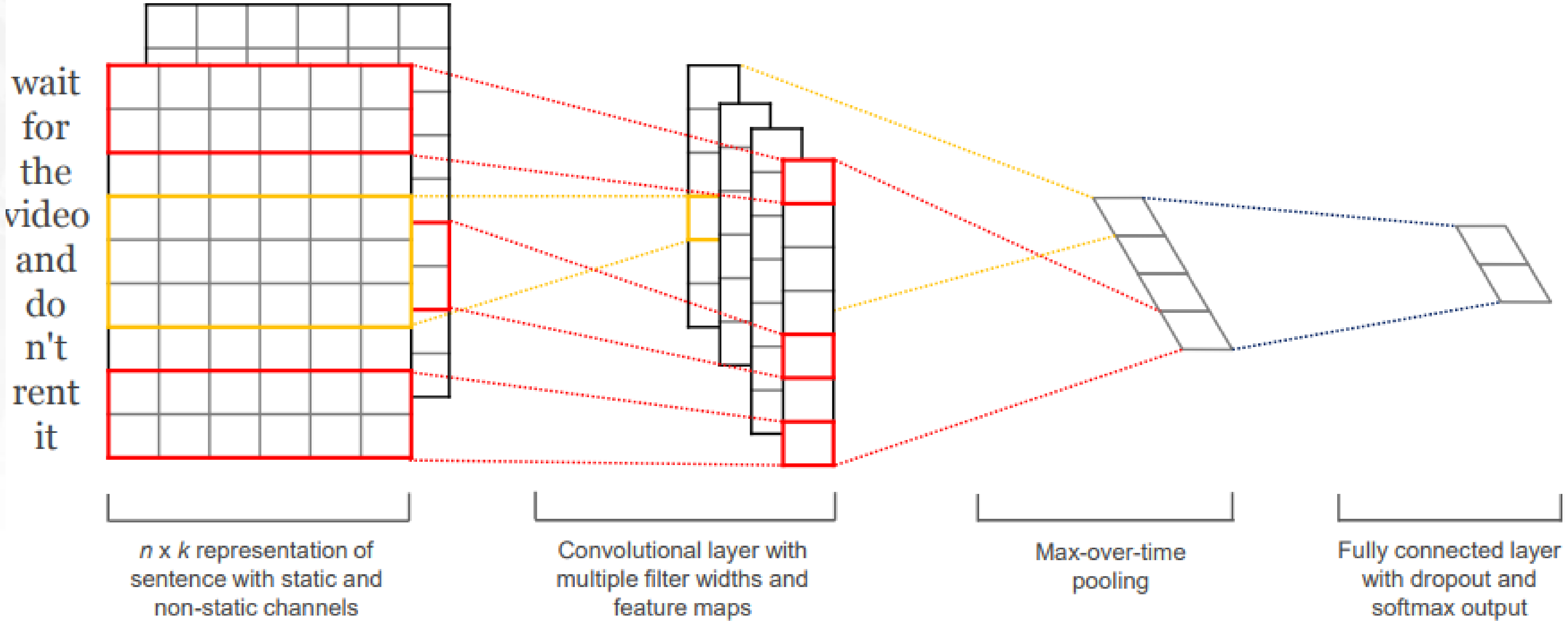


Diamond
Please Contact
133255[redacted]840

www.163.c0m
50USD for ultimate weapons

&$!#%

Text Classification



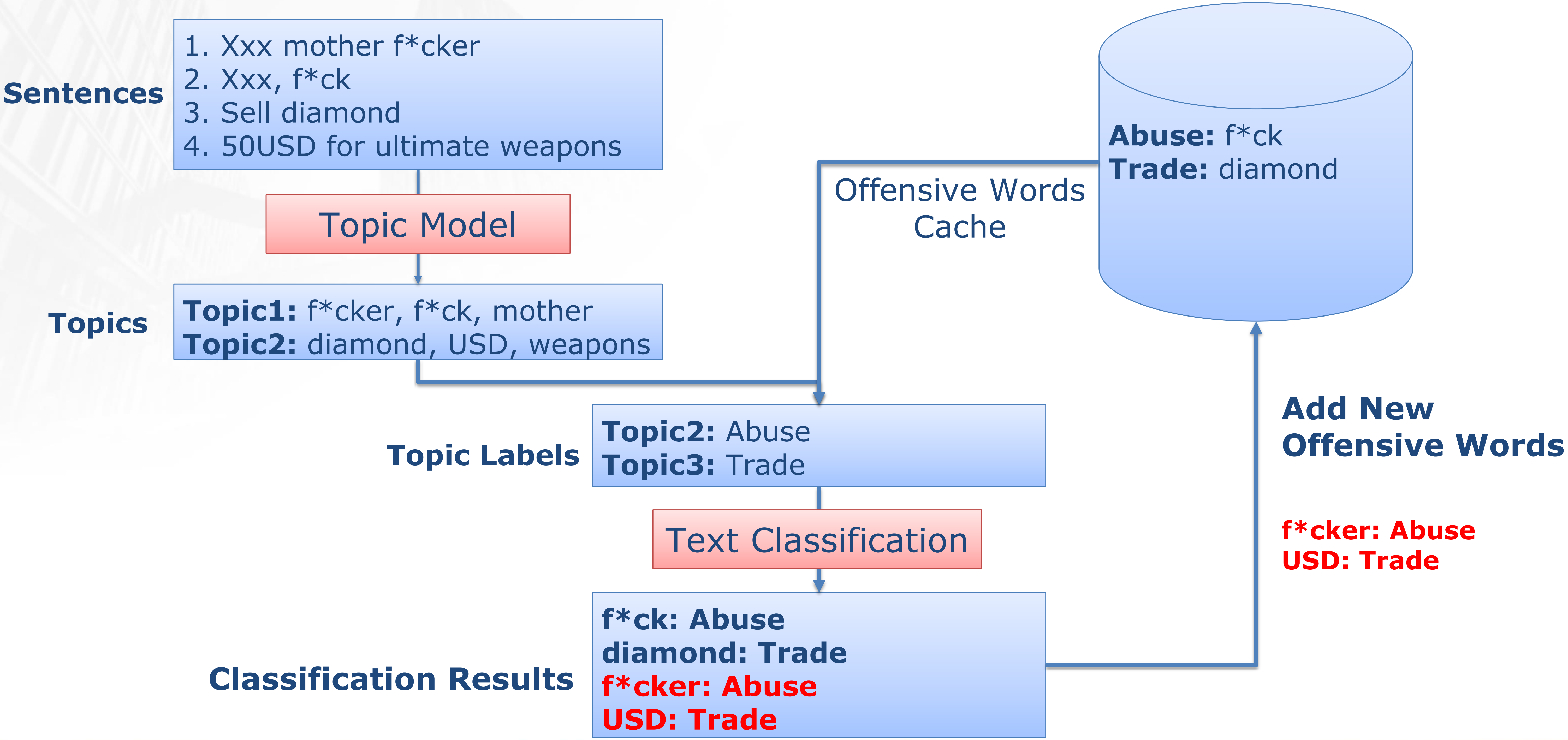
wait
for
the
video
and
do
n't
rent
it
n x *k* representation of sentence with static and non-static channels
Convolutional layer with multiple filter widths and feature maps
Max-over-time pooling
Fully connected layer with dropout and softmax output

Text CNN

- Use **multiple filters** with varying window sizes to obtain **multiple features**
- Quite simple
- High inference speed

[1] Chen Y. Convolutional neural network for sentence classification.

Cache Mechanism



Sentences

1. Xxx mother f*cker
2. Xxx, f*ck
3. Sell diamond
4. 50USD for ultimate weapons

Topic Model

Topics

Topic1: f*cker, f*ck, mother
Topic2: diamond, USD, weapons

Offensive Words Cache

Abuse: f*ck
Trade: diamond

Topic Labels

Topic2: Abuse
Topic3: Trade

Text Classification

Classification Results

f*ck: Abuse
diamond: Trade
f*cker: Abuse
USD: Trade

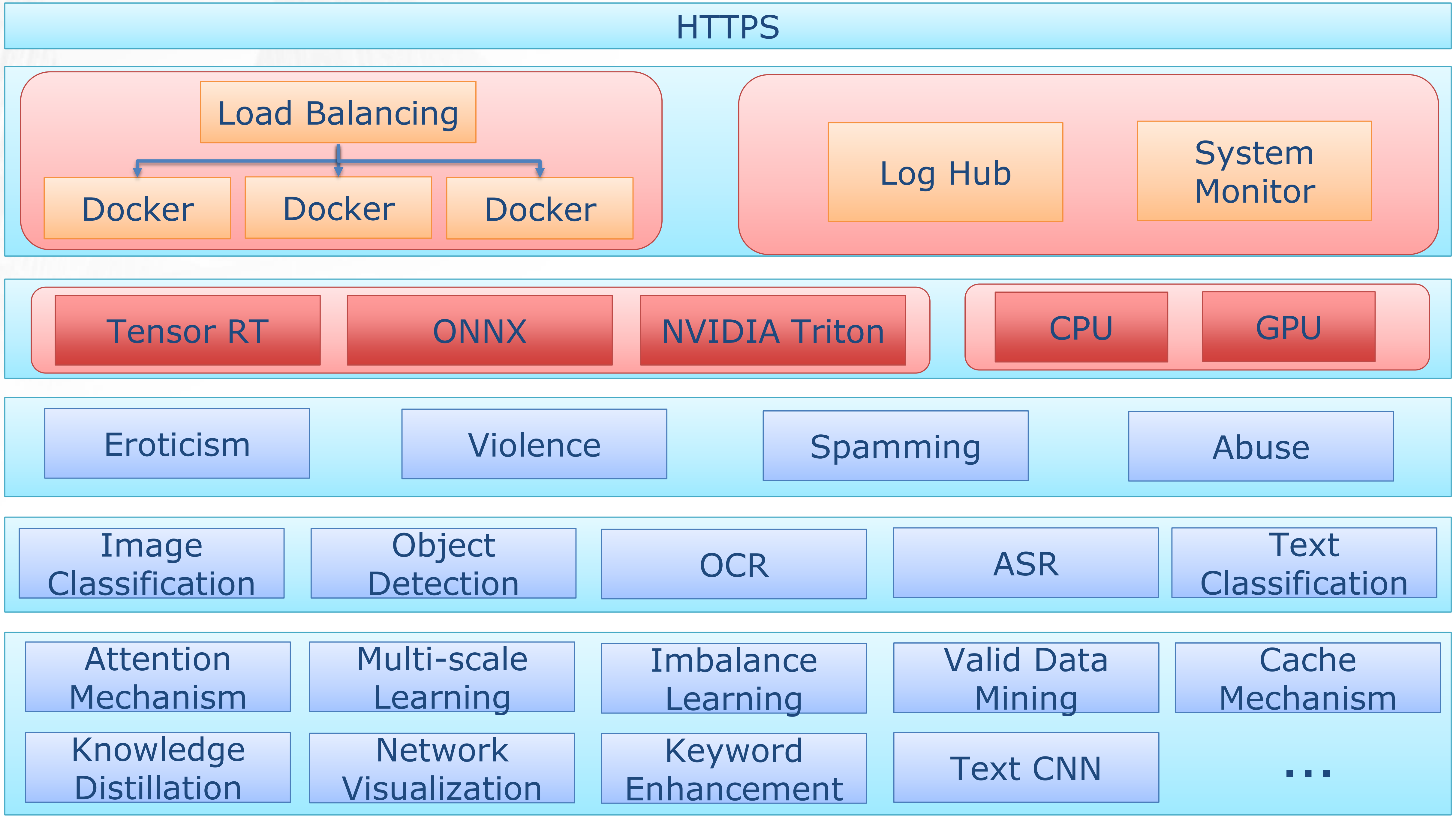
Add New Offensive Words

f*cker: Abuse
USD: Trade

Outline

- Introduction
- Image Sub-system
- Audio Sub-system
- Text Sub-system
- **Application**

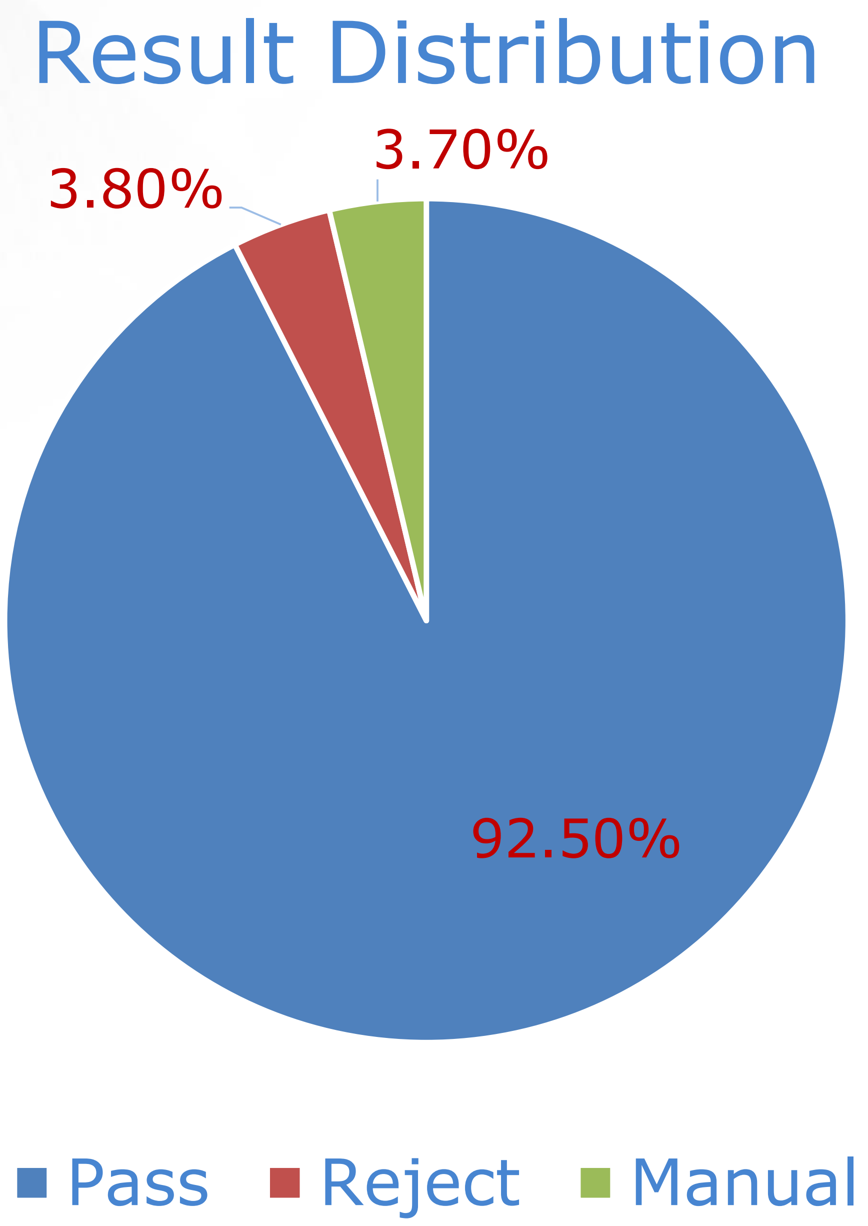
Architecture



Application

- **High precision: 98%** for image and **90%** for audio
- **High performance:** In total, more than **3,200,000,000** images and **73,000,000** hours of audio data are processed
- Has been running **stably for several years**
- Has used in **almost all products** of NetEase Games

Result Distribution



Takeaways

- When build a multi-modal moderation system
  - **Data is critical.** We have introduced the methods for data collection, cleaning, mining and improvement.
  - **Model update and optimization never stop**. Some of the methods we adopted are shown and can be used as a reference.

THANKS FOR WATCHING

China: https://hr.game.163.com/recruit.html

Overseas: https://www.neteasegames.com/careers

Twitter Facebook Youtube