

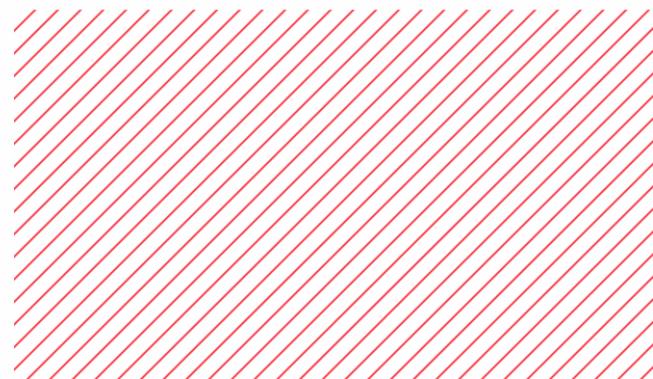
академия
больших
данных



ML System design

Эдуард Тяントов

Руководитель направления машинного обучения
в Почте и Портале



Plan

- Steps of ML System Design
- 2 Case studies
 - OCR @Scale
 - Face recognition on video

Steps of System Design



Steps

1. Problem statement
2. Data
3. MVP
4. Model
5. Metrics
6. Deployment
7. System design



1. Problem statement

1. Clear objective
 1. Business, Error budget
 2. ML, FP vs FN
2. Scale
3. Inference budget & other restrictions

2. Data

1. Data sources
2. Supervision signal
3. Feedback loop

3. MVP

Goal – to learn as much as possible

1. Viability
2. Open questions
3. Potential problems/risks
4. New restrictions

4. Model

- Feature engineering
 - connections between different types of data
 - hand-crafted/automated/statistics
- Model
 - Restrictions
 - Simplicity
 - Explainability
 - ...
- Architecture, losses, training, ...

5. Metrics



Specify target metrics

6. Deployment

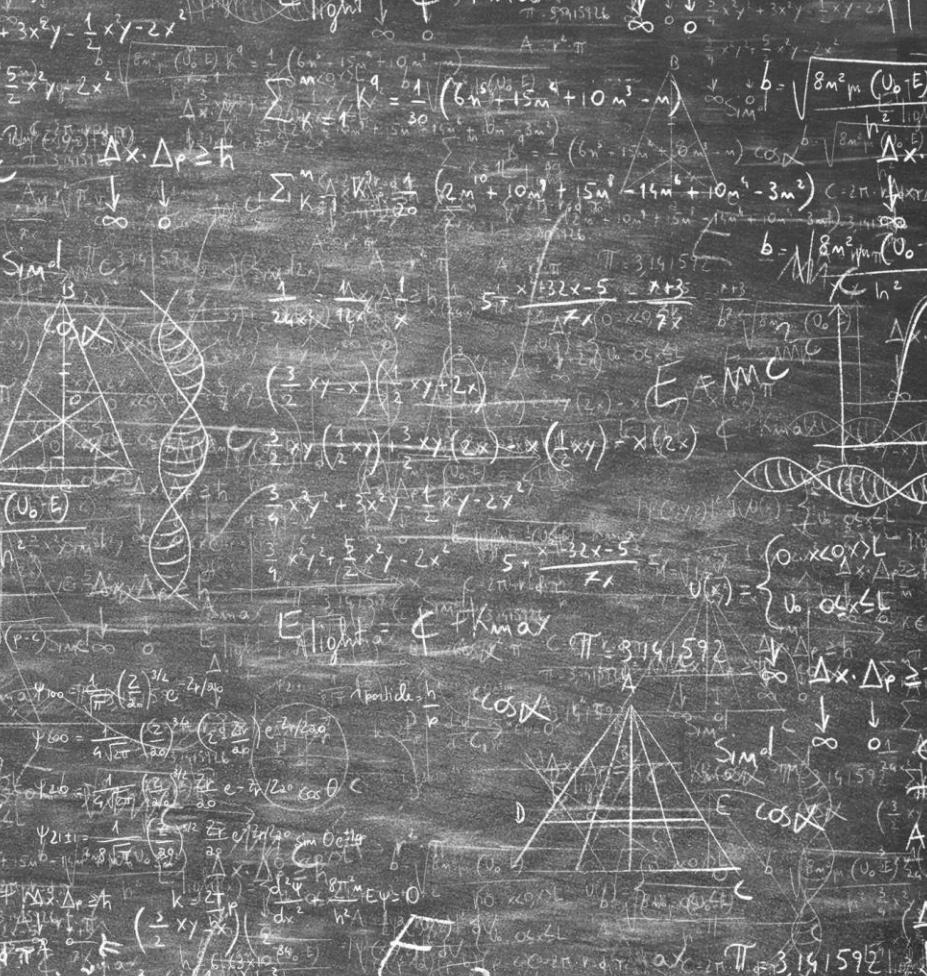
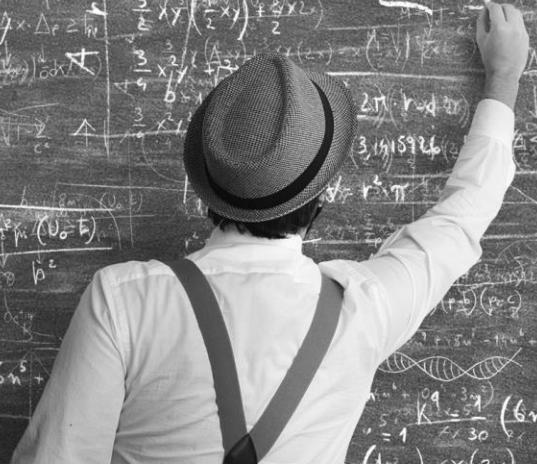
- A/B testing
- Monitoring
- Framework

7. System design*

Infrastructure around the ML component:

- Databases
- Scalability
- Availability
- Security
- Trade-offs

Case study: OCR



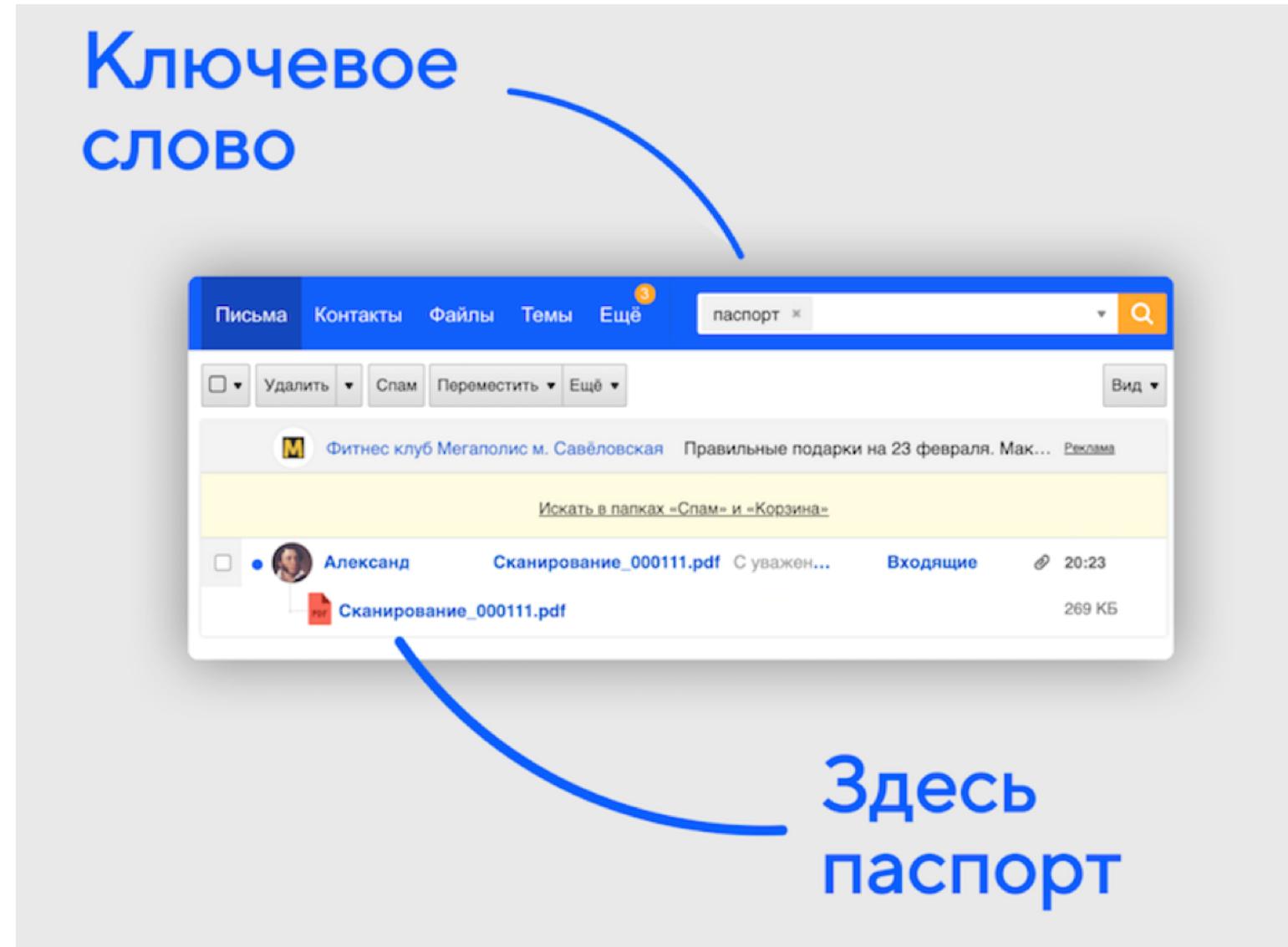
Application: Antispam

- Spam using Images/PDF
- Adaptation/Adversarial

Здравствуйте! Меня зовут Ирина.
Приглашаю на надомную работу. Работа в
интернете. Сотрудничество с крупной
международной компанией. Законно. З/пл -
через банк, стаж, соц. отчисления.
Можно совмещать с основной работой.
Хотите узнать больше о проекте???
Обращайтесь в личку с пометкой
«Кандидат».

Application: Email service

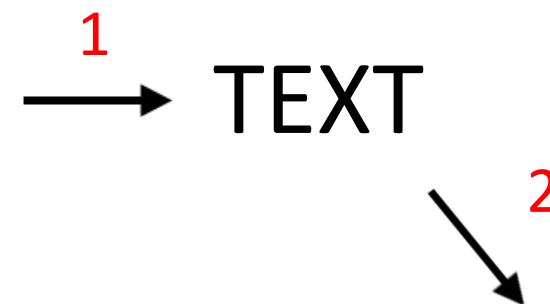
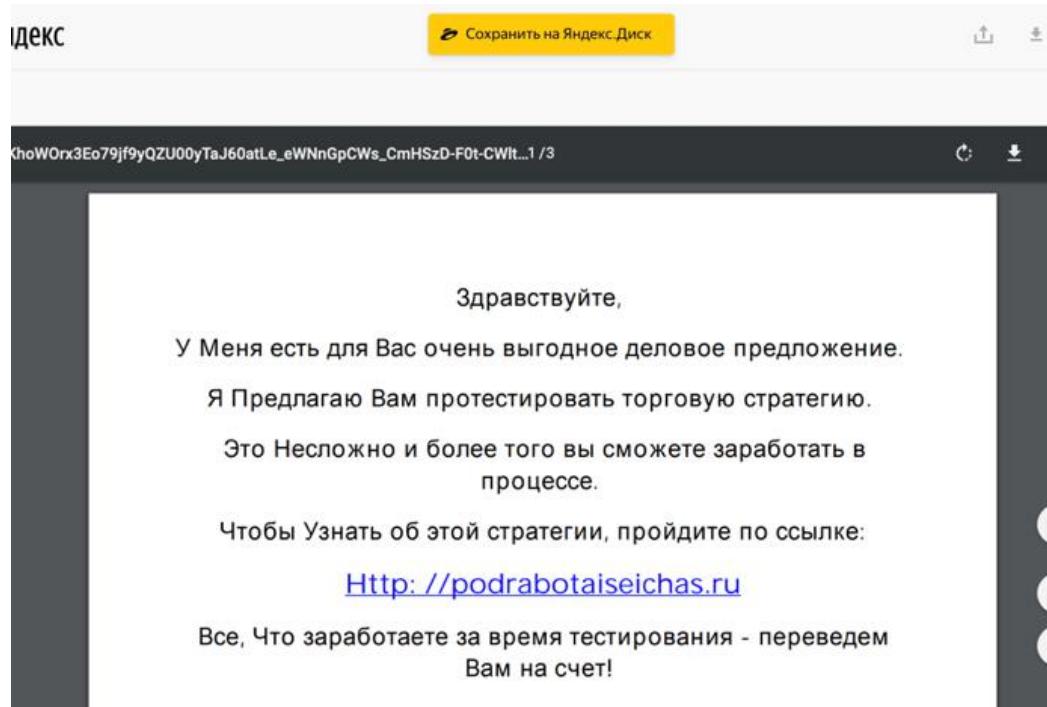
- Text search
- Document types



1. Problem statement

Objective

- Embedded component: for other ML or search engine

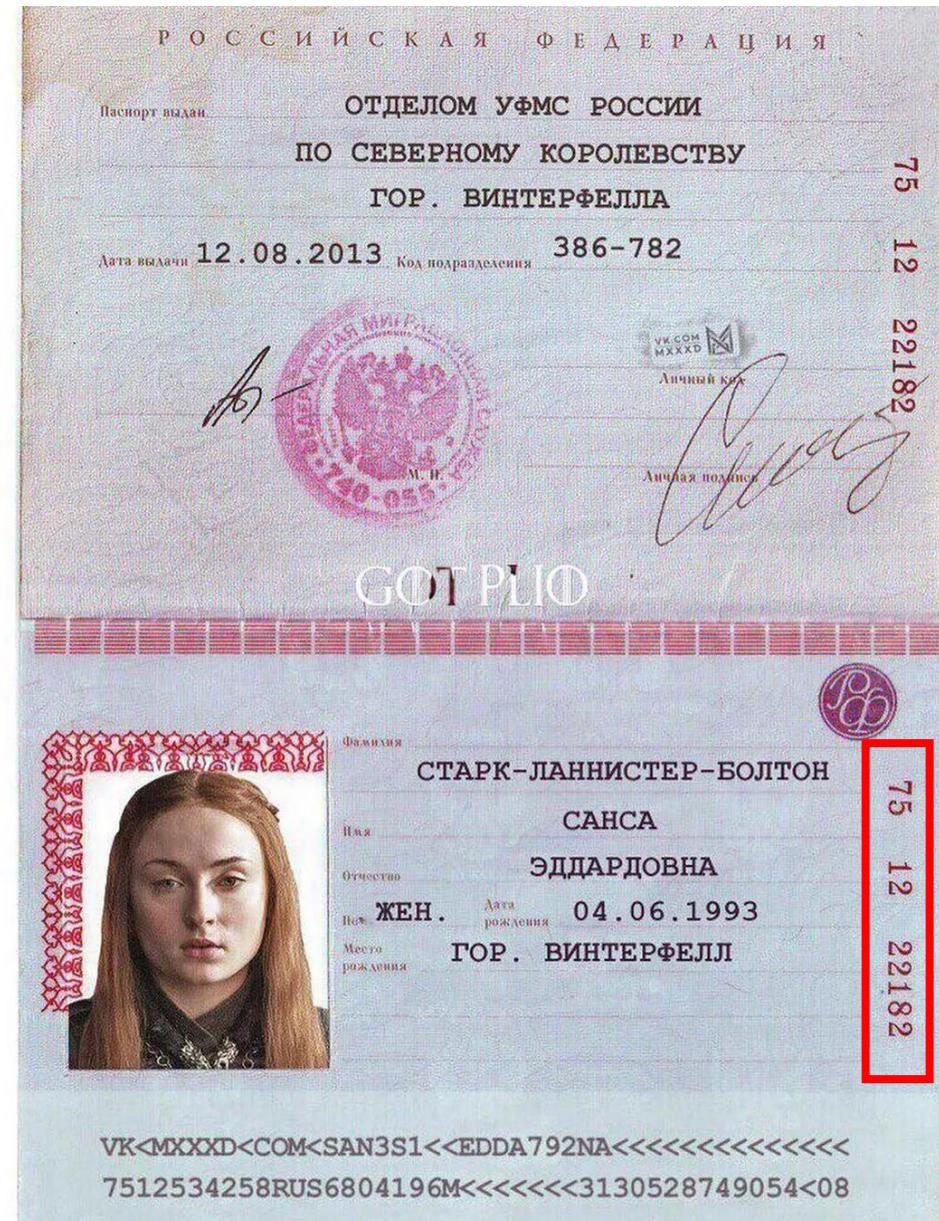


Spam
Classifier

Embedding

Objective

- Errors are not critical
 - != passport recognition
 - which type of error more important ?



Objective

- Multiple languages
- Decent accuracy: Word Error Rate (>98%)
- Adaptability

On VIM
waving a wand, But if
you could still try per
have to really know W
understand some ethi

еbe сюрприз и подарить! Получить можешь (<http://www.salonskincares.com>)
ривет! Ну, как с чайкомто? напомнил он, потирая РУКИ
Info:
Scholes. Sue cont
<https://www.salonskincares.co.u> <https://www.salonskincares.com>
alge Peterson.pdf
6604 Sberbank.ru <<http://lk.mailpost.ru/track/r> http://www.yahoo.com Номер транзакции: отправитель:

T ANFORDERN >><http://asset.m09/ia>
Body: Добрый день A SVBTz+u ЗН
oo.com Номер транзакции: 37673831 Отп
stainabilityrussia.ru Заполнена web-форма
[https://www.google-analytics.c](https://i.emlfiles.com/themeit>Hello Вам нач
<a href=) Весна в
ботки только для тебя =>> заходи скорее Ут
Доброе утро! Зачем было нас туда посыпало
your ... you can reset it by visiting <http://>
ума, в какой-то степени 74/человека ориентир
Салют! Он дрожал головы до ноги в обиде, за

Scale & Inference budget

- 50,000 rpm
- <100ms on CPU+GPU for Antispam
- <500ms for Email/other applications

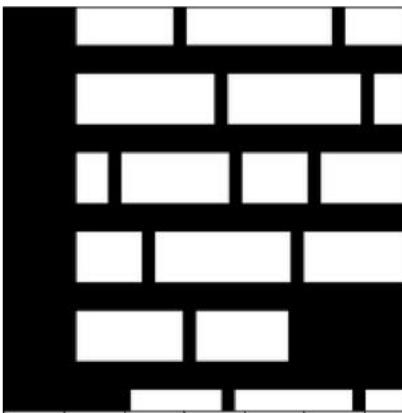
2. Data

Sources, Supervision

- Labelling is very hard
 - Doable: hard examples for recognition

even though par
"That's awful! Y
"I think the wor
the wrong quest
harm than

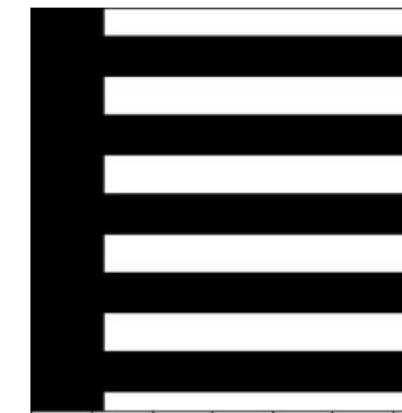
image



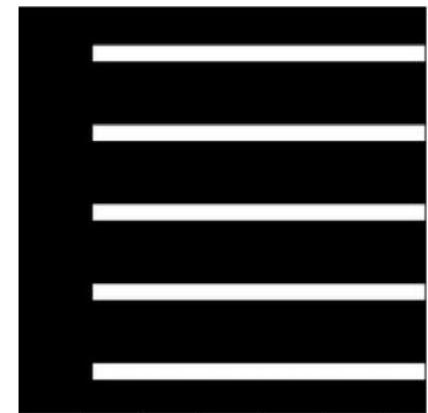
words



spaces

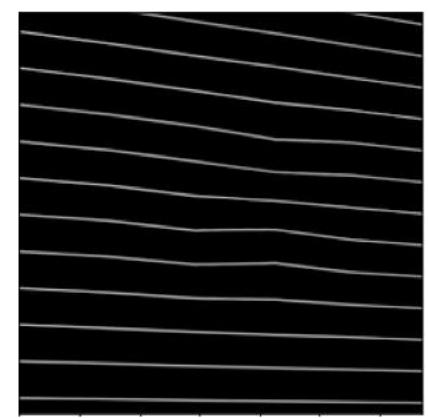
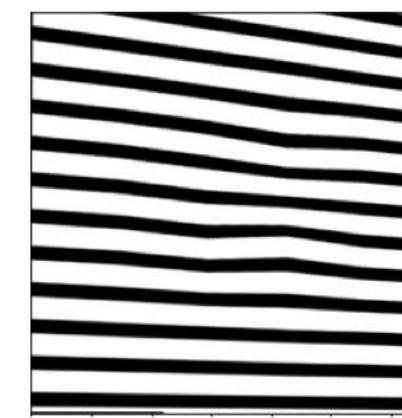
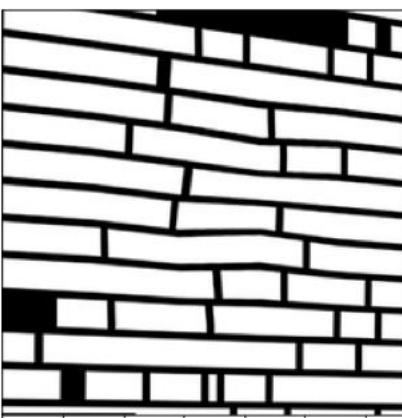


rows



delimiters

86 1
Registrierung auf mario. Ihr Ko
442641998 Sberbank.ru<https:
лайуъщлт текже счимыгщф
сийнэчф С.2№ЧсБ9 бжу лдмз
78686275193 SBERBANK.RU<
Сообщение! Дарья Елистратова прислал
Дарья Елистратова прислал
Приветствую Вас! Через ЧА
они могут ожидать от ме
Hi CobrynGoophorry, This notic
Ads Buy Sell | UK Marketpla
es



Data: generator

- Opensource Backgrounds
- Texts: literature, spam, documents



Data generator: example



Pick a background

{cid:49c9401fdfffe0e092d372f79@kacsp.
<https://www.paradies-bettensho> <https://www.yahoocom> Номер транзакции: 45925357
<https://www.yahoocom> Номер транзакции: 14321671
0438898819 Sberbank.ru<<https://drive.google.com> 8734205 Производим выплаты
Совет от друга Ваш рекомендует ко
т Типат и ск Request ONel@jKuEQ Рен
выплаты сегодня lyuba.kustova.81

+Text

www.paradies-bettensho https://www.yahoocom Номер транзакции: 4592535
ahoo.com Номер транзакции: 1432167
438898819 Sberbank.ru<https://drive.google.com> 8734205 Производим выплаты
Совет от друга Ваш рекомендует ко
т Типат и ск Request ONel@jKuEQ Рен
выплата сегодня lyuba.kustova.81

Augment

Data: Feedback loop

- No direct feedback ;(
- Adaptation through hard examples
 - how to **mine** them? (see metrics section)

3. MVP

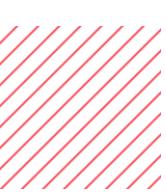


MVP

Heuristics, opensource?

- Tesseract doesn't fulfill neither requirement
 - Inference: CPU, 1-3s
 - Adaptability: bad





MVP

The only option: deploy first prototype CNN

What can we learn?

- Risks: inference budget
- Data: longtail, edge cases

New information: data & backend

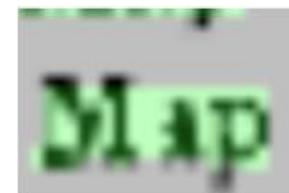
@ mail.ru
group

- Not by images alone – PDF
 - → one more backend + near-online
- Large images (>10mb)
 - → significant strain on CPU
 - → to handle all images – near-online



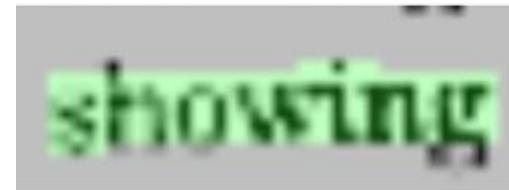
New information: data

1. Language mismatch



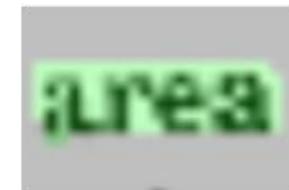
Map

→ “Map”



showing

→ “showing”



area

→ “area”

New information: data

2. Trash text



“ЭЩКЕРЕ”

New information: data

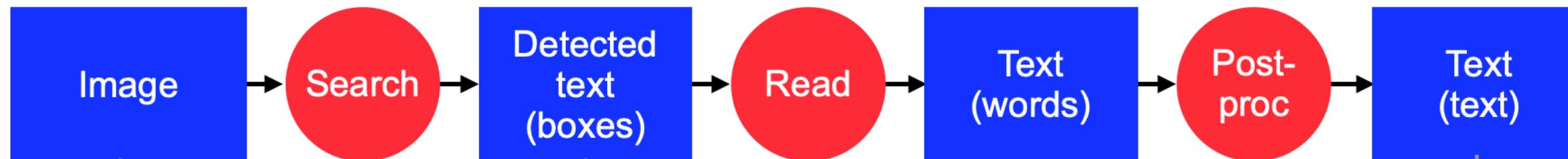
3. “misprint” errors

TABLE 6. → “TabId 6.”

FIGURE → “Eigure”

4. Model

Model



Data Fest⁶



Data Fest⁶

{'Data',
'Fest6'}

Data Fest6



How to handle edge cases?

@ mail.ru
group

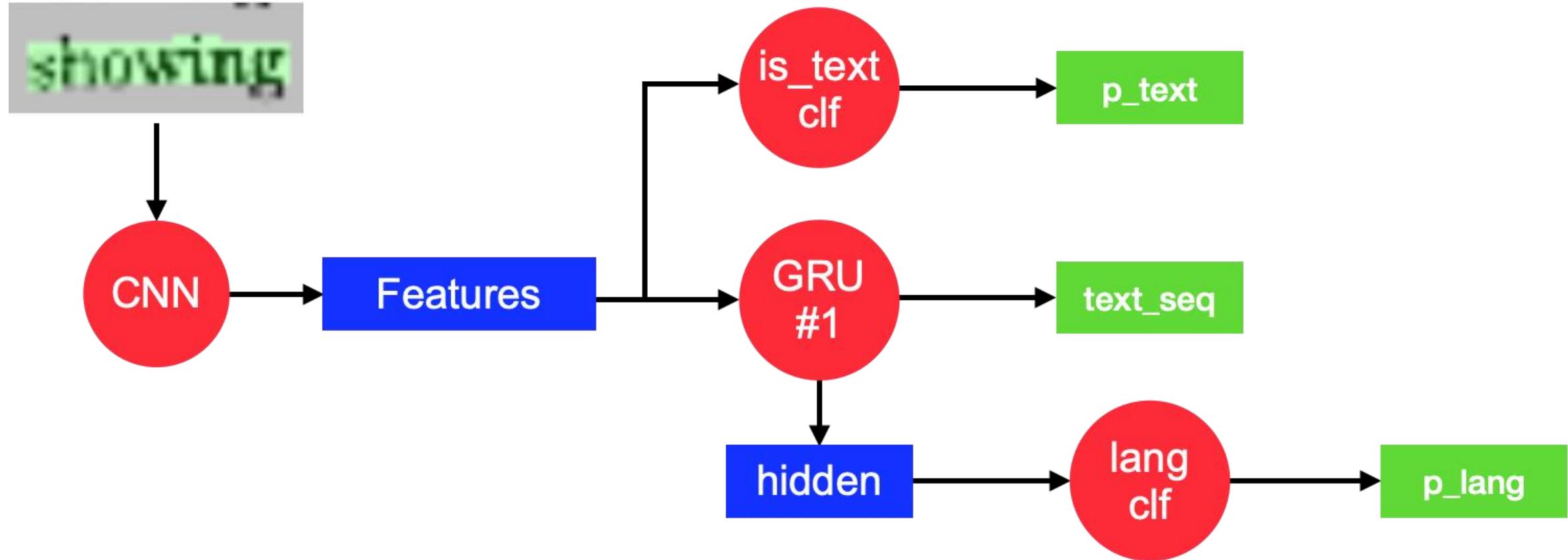
1. Language mismatch
2. Trash text
3. “Misprint” errors

Edge cases

- Simple heuristics > ML
- ML > complex heuristics



Multi-head for Recognition



Misprints

How to tackle:

- Word model
- One more head
 - word, BPE, n-grams, ...

TABLE 6.
FIGURE

→ “Table 6.”

→ “Figure”

5. Metrics

5. Metrics



1. Complaints on images (antispam)
2. Added value to Classifier's metrics
 - Average BPE length
3. Accuracy

BPE metric usage

We can use BPE to estimate trash text probability

“Hello, Made students” – High BPE length

“dsakljds saknx2 sdak99” – Low

Effective for error mining

6. Deployment

A/B testing

- How to split ?
 - A/A test
- Which metrics ?
 - Only responsive (added value, BPE)

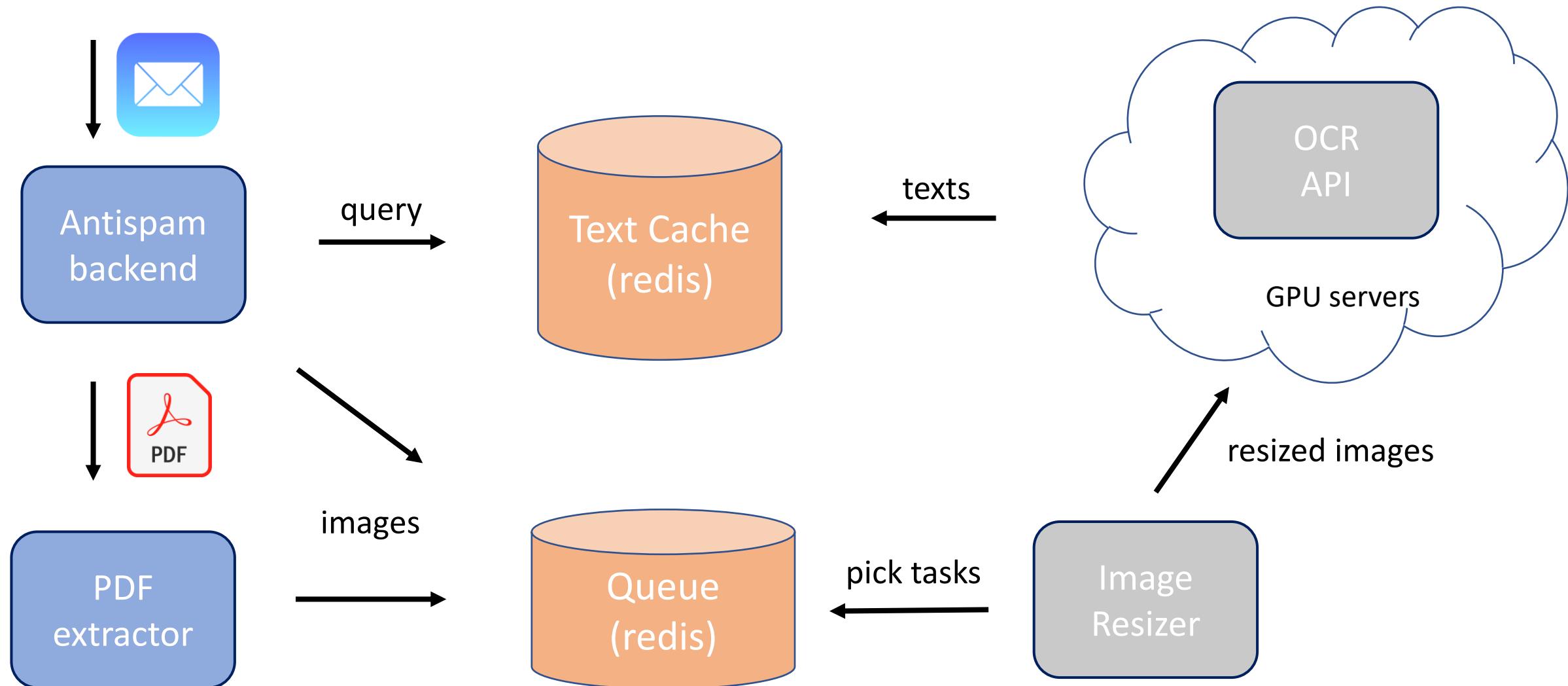
Monitoring

Find error in a pipeline as soon as possible

- Integration tests
- Data tests
- Monitor metrics
 - Technical: volume, response time, ...
 - Product

System design

Antispam design: near-online



Emails with first/unique images/PDFs are post-processed

Case study: Face recognition on video



FR on video

Imagine we've already have FR on photo
(see “Metric learning” lecture)

Video > independent sequence of photos

Application on site

- Access control
(known people)
- Track intruders
(unknown)



1. Problem statement

Objective

FP vs FN ?

- Access control → very low FPR < 0.01%
- Recall (rank1) > 95%
 - How to initialize the system ?

Restrictions

- Scale: N cameras
- Inference budget: < 1s
 - people don't like to wait



2. Data

Data Sources

Specific hardware (cameras)

→ Tuning on “home” installation

No feedback loop ;(

3. MVP



MVP

@ mail.ru
group

Video = sequence of photos ;)

What can we learn?

- Cameras: blurry, bad angles, too far, low resolution
- How many extra persons ?

New info: blur

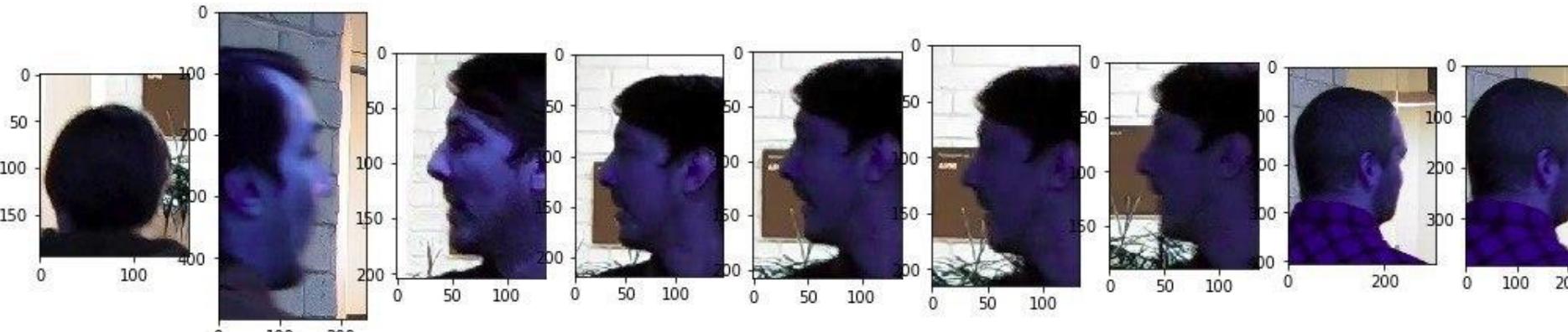


New info: Occlusion

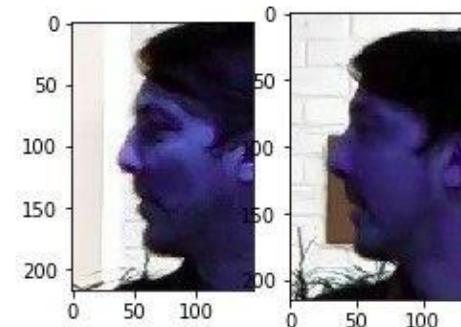
@ mail.ru
group



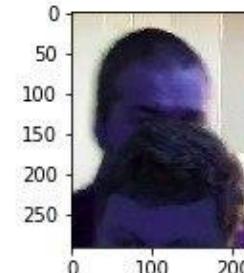
- Unbelievable amount of
 - Edge cases
 - “Splits”



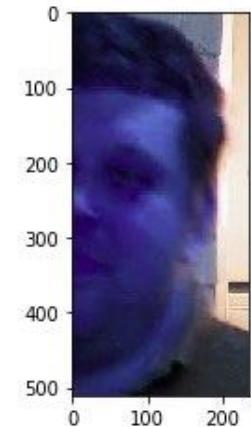
person16--vision102
(0.988, -0.7221, nan)
(0.9977, -0.7159, nan)



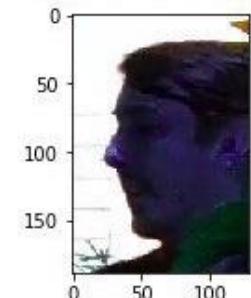
person6--vision102
(0.3932, 0.8373, nan)



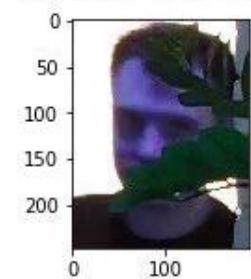
person13--vision102
(0.4805, 0.7056, nan)



person15--vision102
(0.988, -0.8557, nan)



person8--vision102
(0.5895, 0.7479, nan)



Cameras

- Bad tuning
- Weird angles

→ tests & guides

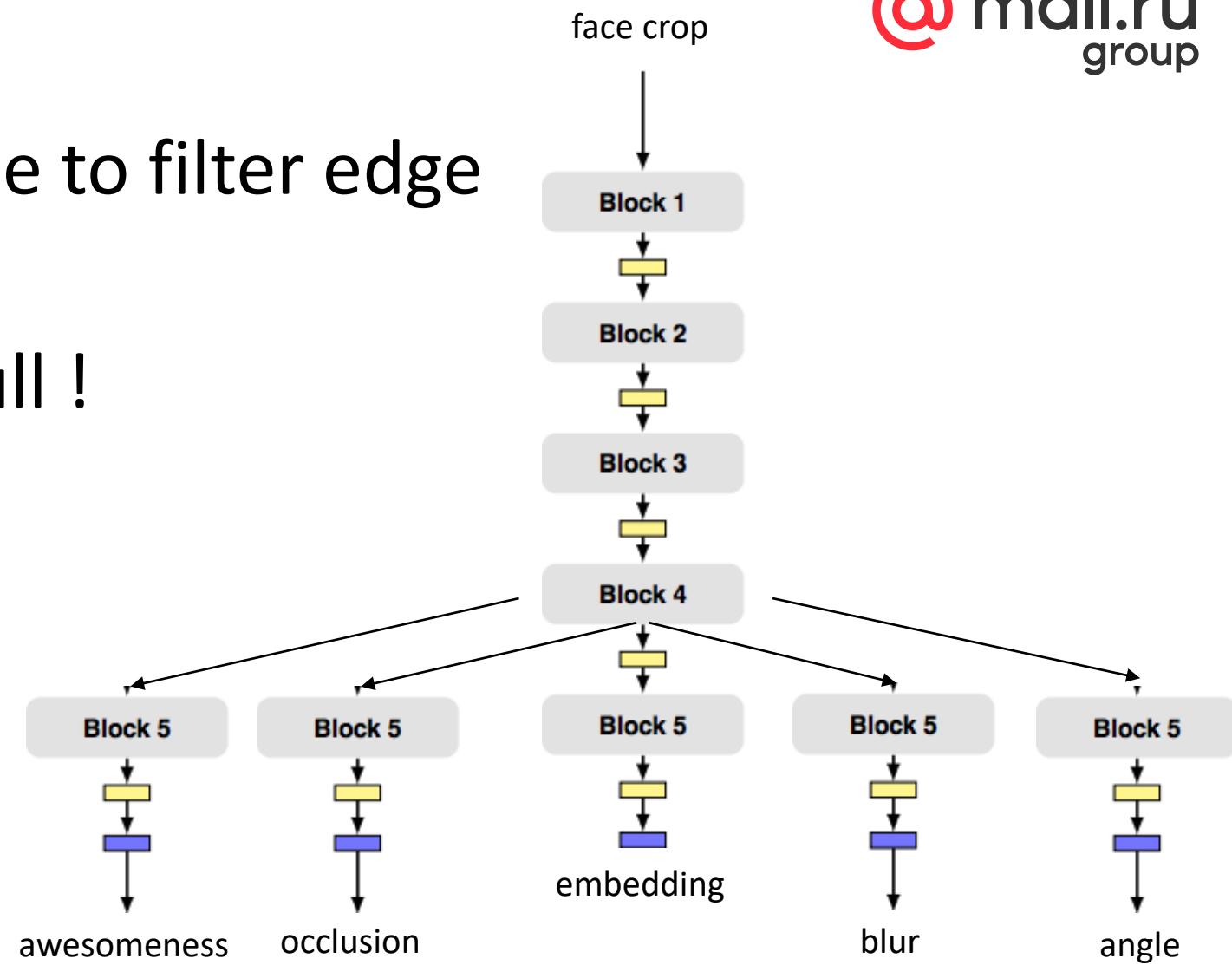


4. Model

Filtering

Predict every aspect of face to filter edge cases

Gives us many levers to pull !



Tracking

- Use sequence!
- Track instead of face



13.03.2020 12:05:05
Фронтальность: 0.8895.
кадр: 9186518
[Распознавание лиц 1](#)



13.03.2020 12:05:05
Фронтальность: 0.9538.
кадр: 9186517
[Распознавание лиц 1](#)



13.03.2020 12:05:05
Фронтальность: 0.9752.
кадр: 9186516
[Распознавание лиц 1](#)



13.03.2020 12:05:05
Фронтальность: 0.964.
кадр: 9186515
[Распознавание лиц 1](#)



Tracking: example



Tracking

Sort/DeepSort uses:

- Bounding boxes
- Embeddings

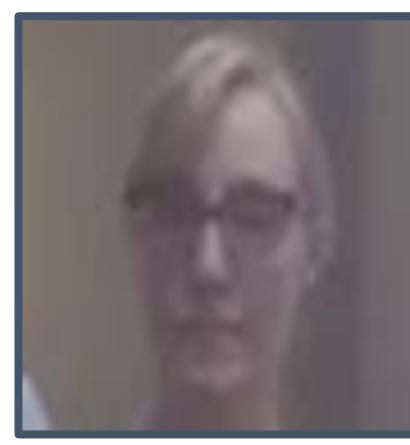


Tracking: best shot

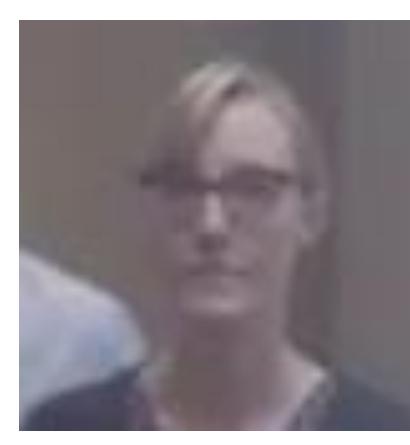
Score = Frontality + Awesomeness + Blurriness + ...



Score: 0.1



Score: 0.7



Score: 0.4



Best

5. Metrics

5. Metrics



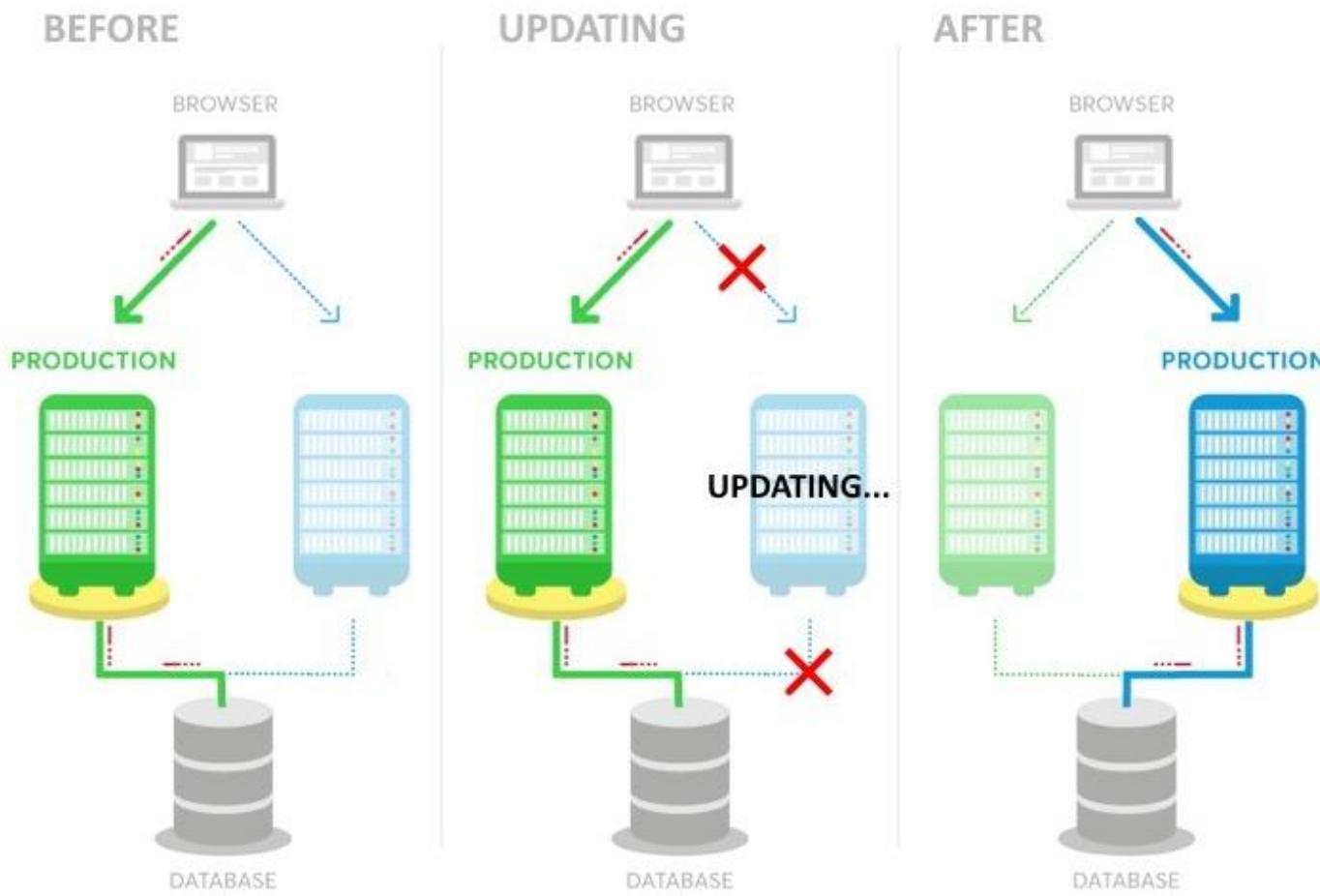
1. FPR, Rank1 (manual)
2. Feedback from users
3. Rank1

6. Deployment

A/B testing

- How we A/B test with new embeddings ?

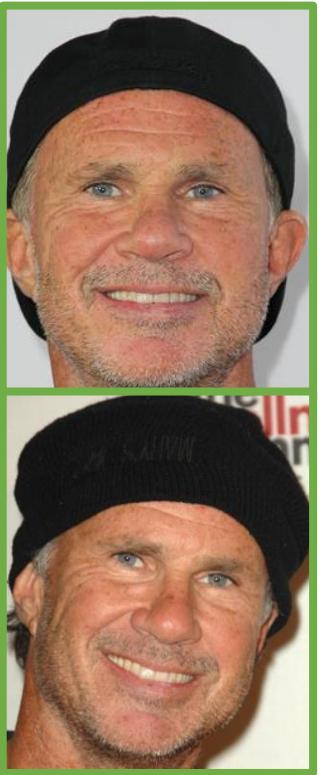
1. Blue/Green deployment



A/B testing

- How we A/B test with new embeddings ?
 1. Blue/Green deployment
 2. **Harmonic** embeddings

Triplet loss: harmonic



Positive

minimize
↔



Anchor

maximize
↔



Negative

$$\text{positive} + \alpha < \text{negative}$$

Triplet loss: harmonic



minimize



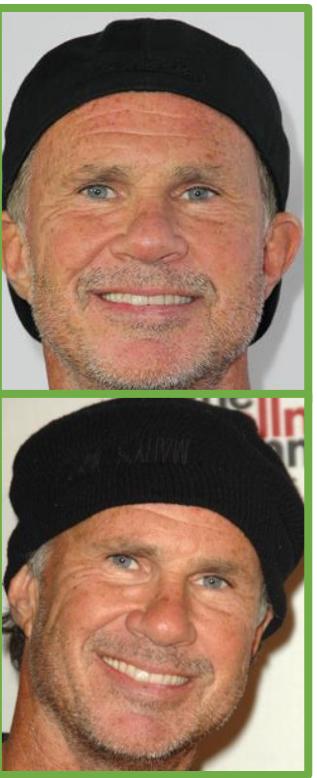
Anchor

maximize

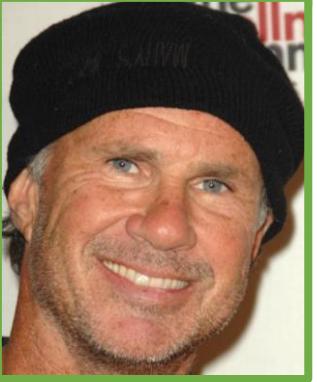


Triplet loss: harmonic

Emb v2



Emb v1



minimize



maximize



Anchor

Mix embedding versions during training

Emb v2



Emb v2



Emb v1



System design

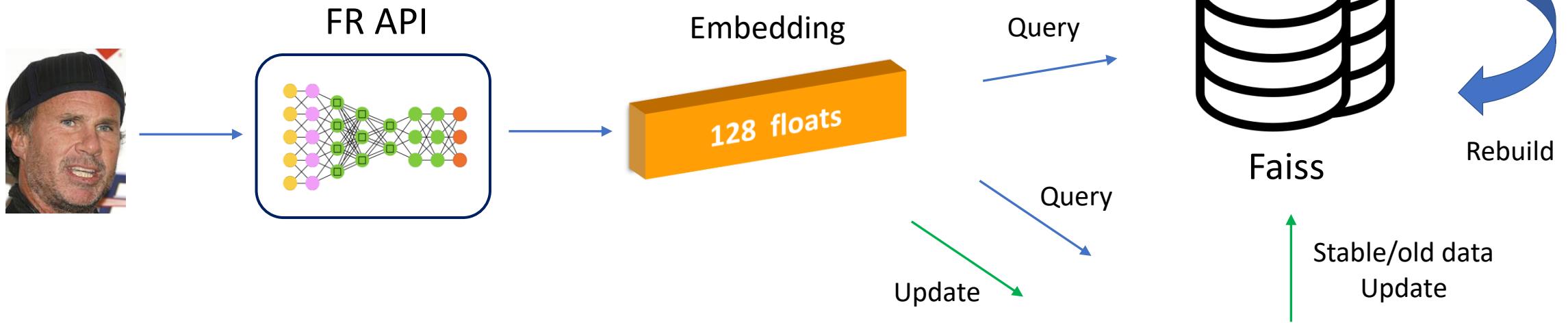
Index

- KNN Index
 - not easy to update it on the fly
 - > Store only stable data
- Online clustering for new

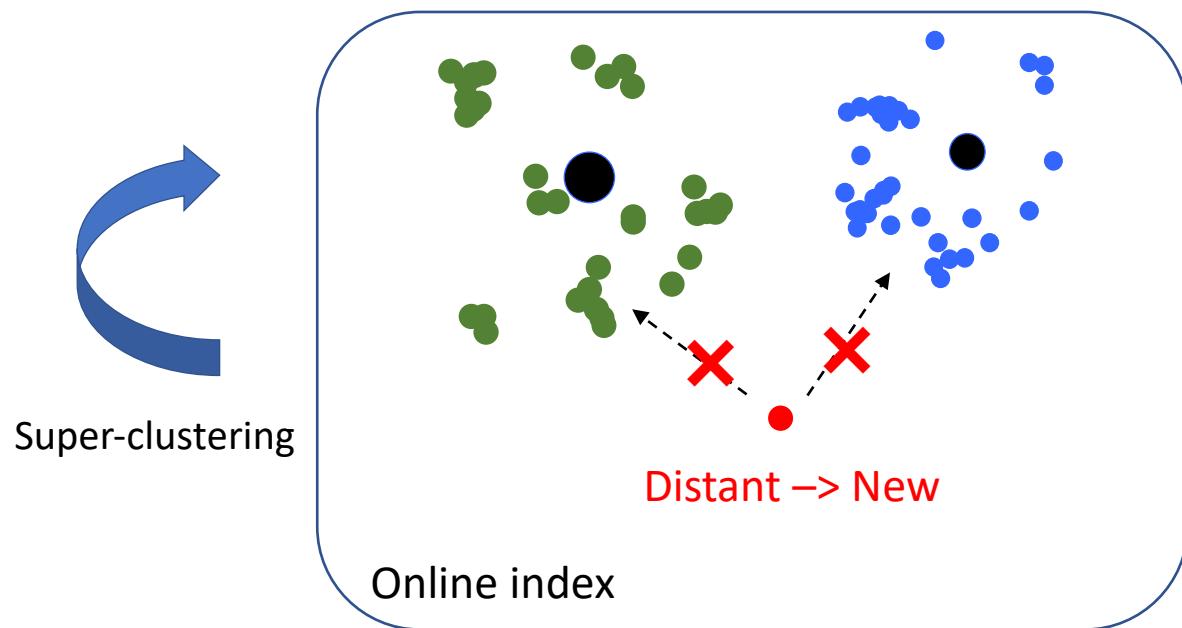


Index (FAISS/Custom)

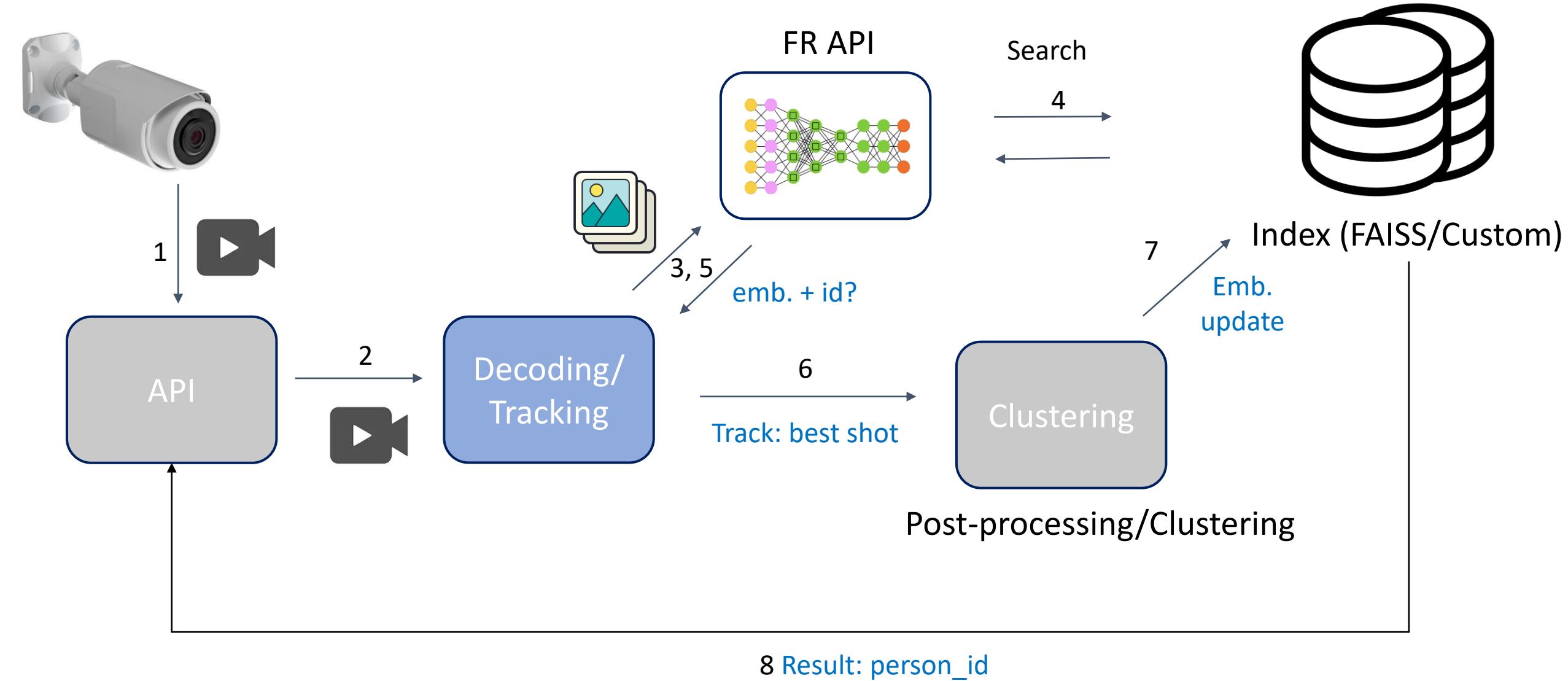
Index



- recent clusters (online)
 - periodic super-clustering
- long-term index (faiss)
 - periodic rebuild



System design



Takeaways



Takeaways

1. Problem statement: Objective, Inference budget
2. Data: Sources, Signal, Feedback Loop
3. MVP
4. Model: Features, Arch, Loss, ...
5. Evaluation metrics (test → prod-ml → product)
6. Deployment: A/B testing, Monitoring
7. System design: Scale, Trade-offs

Контакты



tg: @ed_tyantov



tyantov@corp.mail.ru