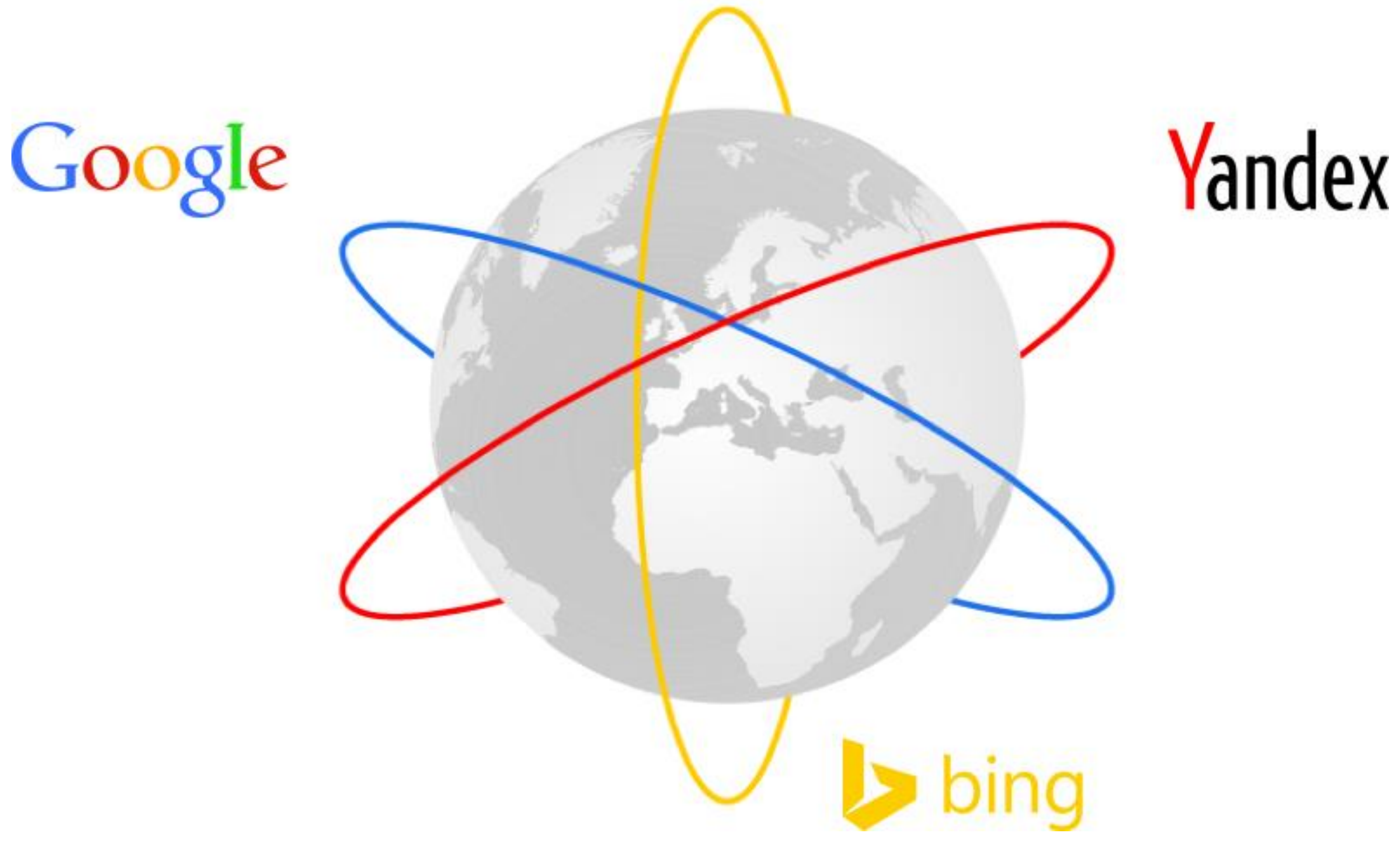
Yandex Data Factory Yandex

SCHOOL OF DATA ANALYSIS

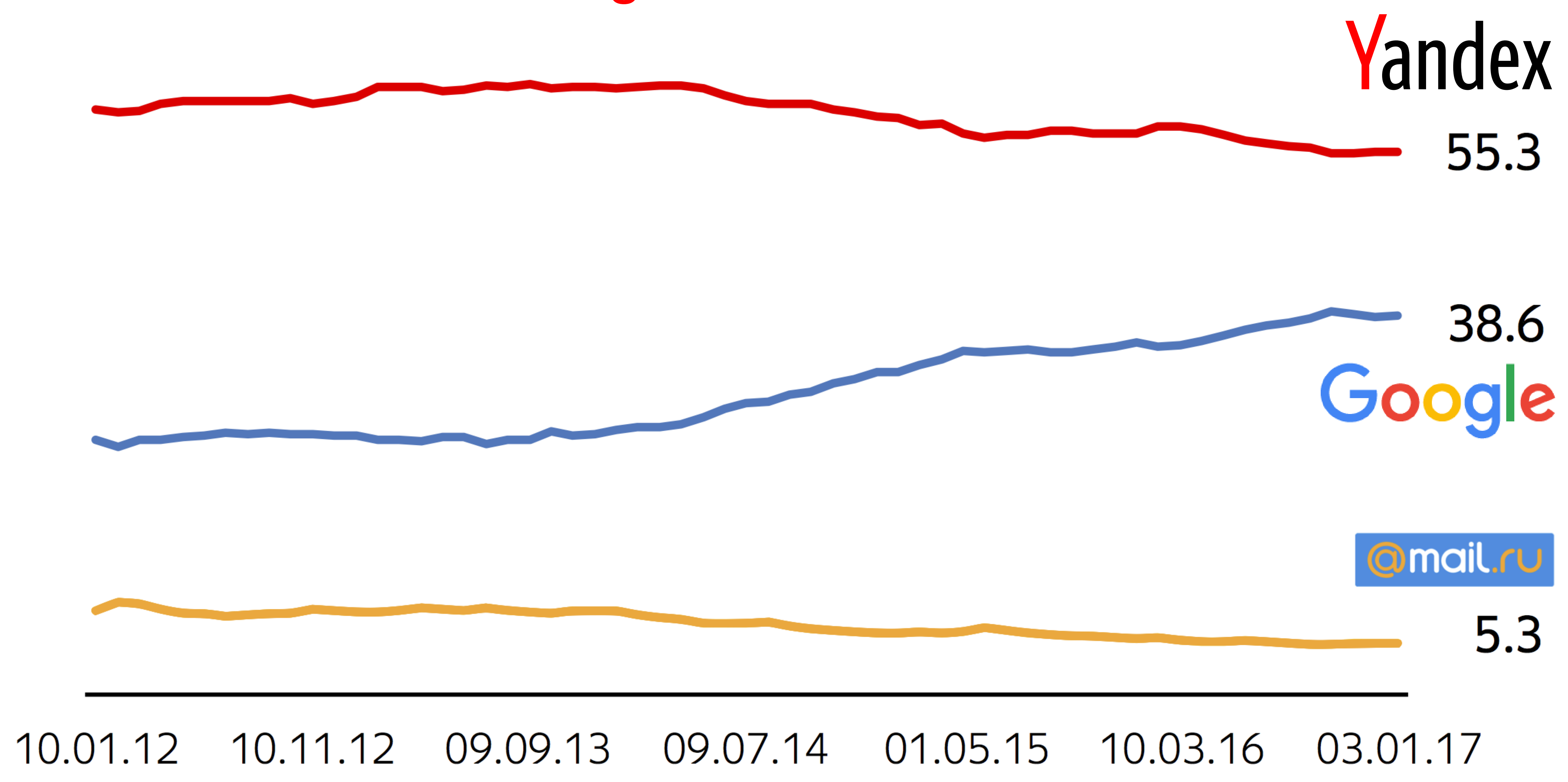
Machine Learning examples at Yandex and beyond

Andrey Ustyuzhanin

Global search engines

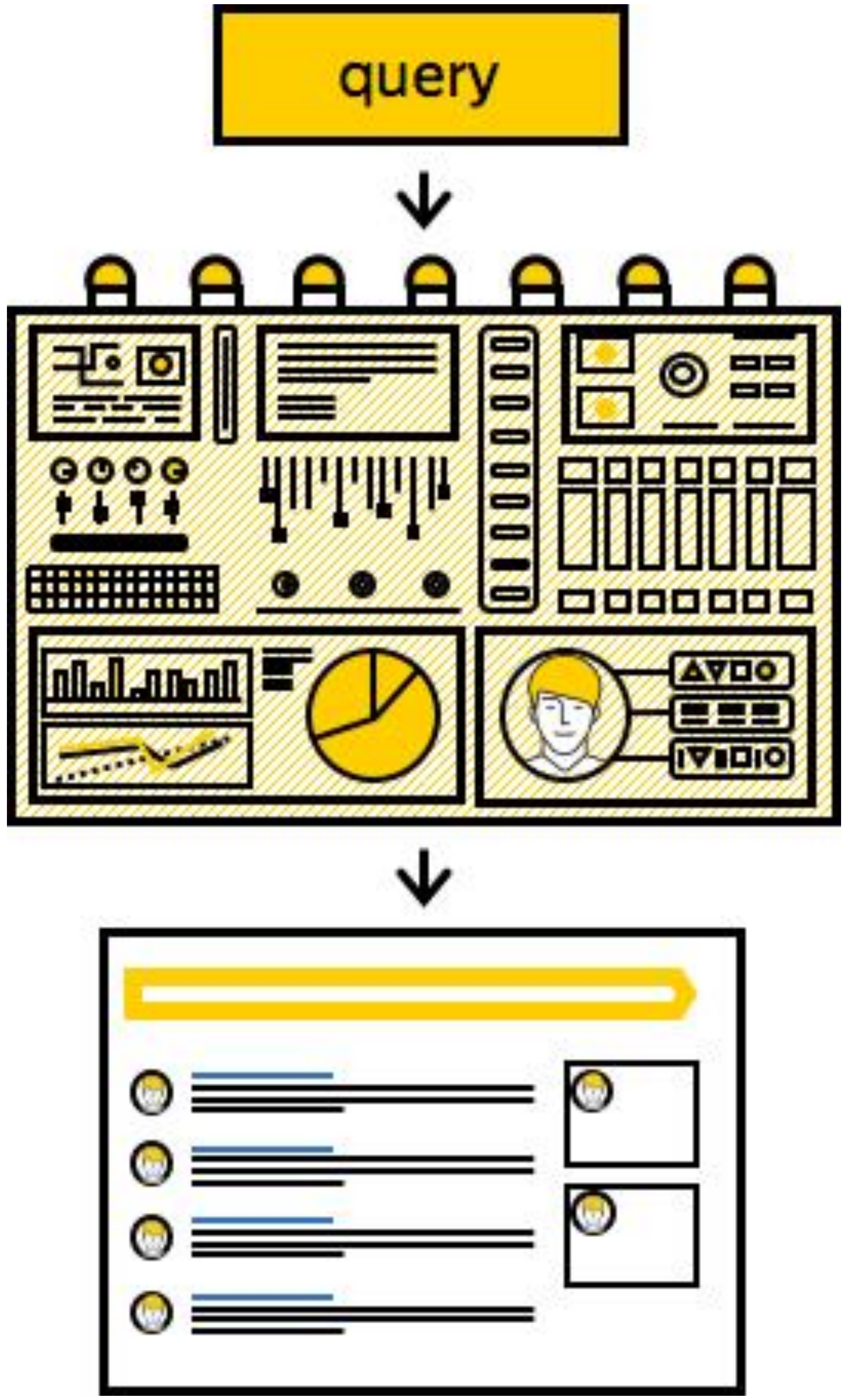


Russian Search Engine Market Share



Source: Liveinternet.ru 2012-December, 2016; includes desktop and mobile

Core Yandex ML task



Search Engine tasks:

- Processing of the entire web (~10^{10} pages, http://www.worldwidewebsize.com)
- Aggregation of users on behaviour analytics and micro-segmentation (~10^8-10^9)
- Personalized delivery of relevant **URLs/Images/Ads** with the highest probability of click

Internal Machine learning application examples:

- Navigation / maps
- Natural Language Translation
- Speech recognition/ synthesis
- Ya.Market (aggregator of Internet Webstores)
- Music streaming / recommendation
- Weather prediction ("nowcasting")
- Ya.Taxi – demand prediction for specific space-time region
- Unwanted content filtering

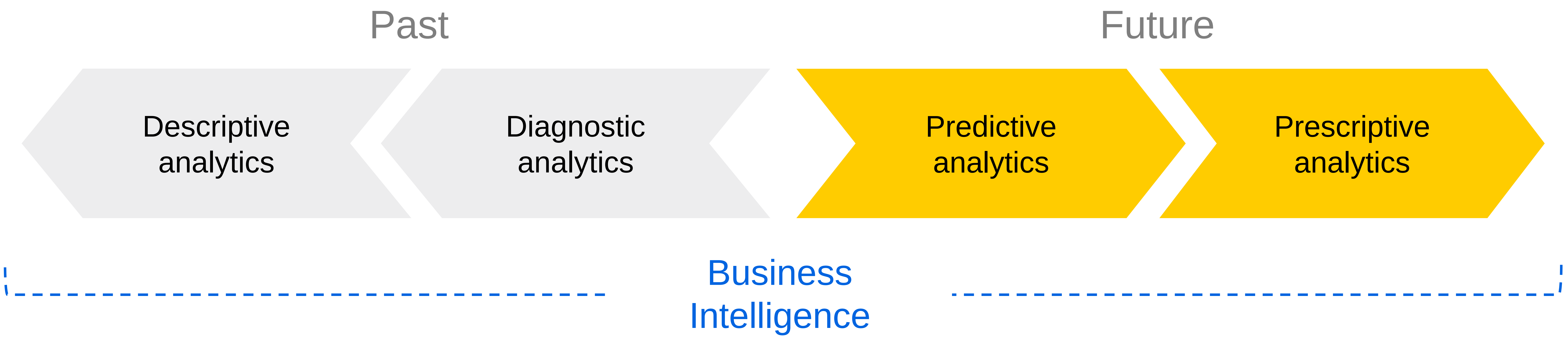
Why ML works

- People
  - Experts, training (http://yandexdataschool.com)
  - Management
  - Product developers
- Tools
  - Diversity of tools
  - Library of pre-trained models
  - Infrastructure to support research & deploy cycle
  - Experiment running & comparison
  - Visualisation & Reporting
- Metrics (you cannot improve what you cannot measure)
  - Well-defined on product-level
  - Proxy metric (ML-metric) for optimisation

SCHOOL OF DATA ANALYSIS

Among our clients and partners:

Focus of our expertise



Key domains of expertise:

- Behaviour analytics
- Time series and anomaly detection
- Geospatial analytics
- Voice recognition and computer vision

Domains of expertise

Behaviour analytics

Data: web-site behaviour logs, clients' profiles, transactions, orders, purchase history, billing data, click-stream, etc.

Solutions: personalised recommendations, "next best offer", churn prediction, loyalty management etc.

Time series and anomaly detection

Data: telemetry data, data from consumer meters, sales volumes, any historical data on events flow

Solutions: fault prediction, predictive maintenance, demand prediction, fraud detection, etc.

Geospatial analytics

Data: geolocation data, transport circulation data, routes, etc.

Solutions: logistics optimisation, road network management, retail network management

Voice recognition

Data: call centre recordings, etc.

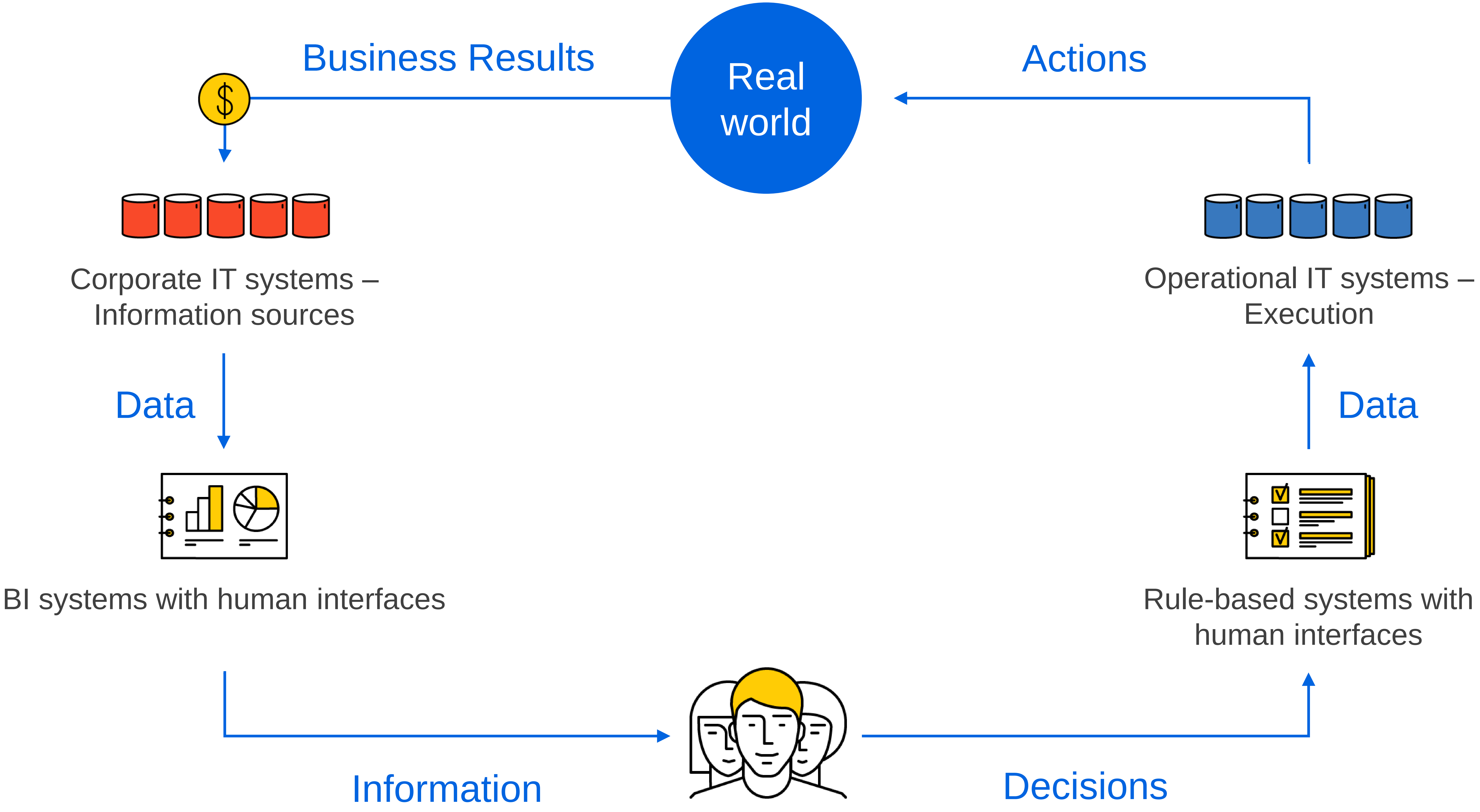
Solutions: personnel screening, script optimisation, etc.

Computer vision

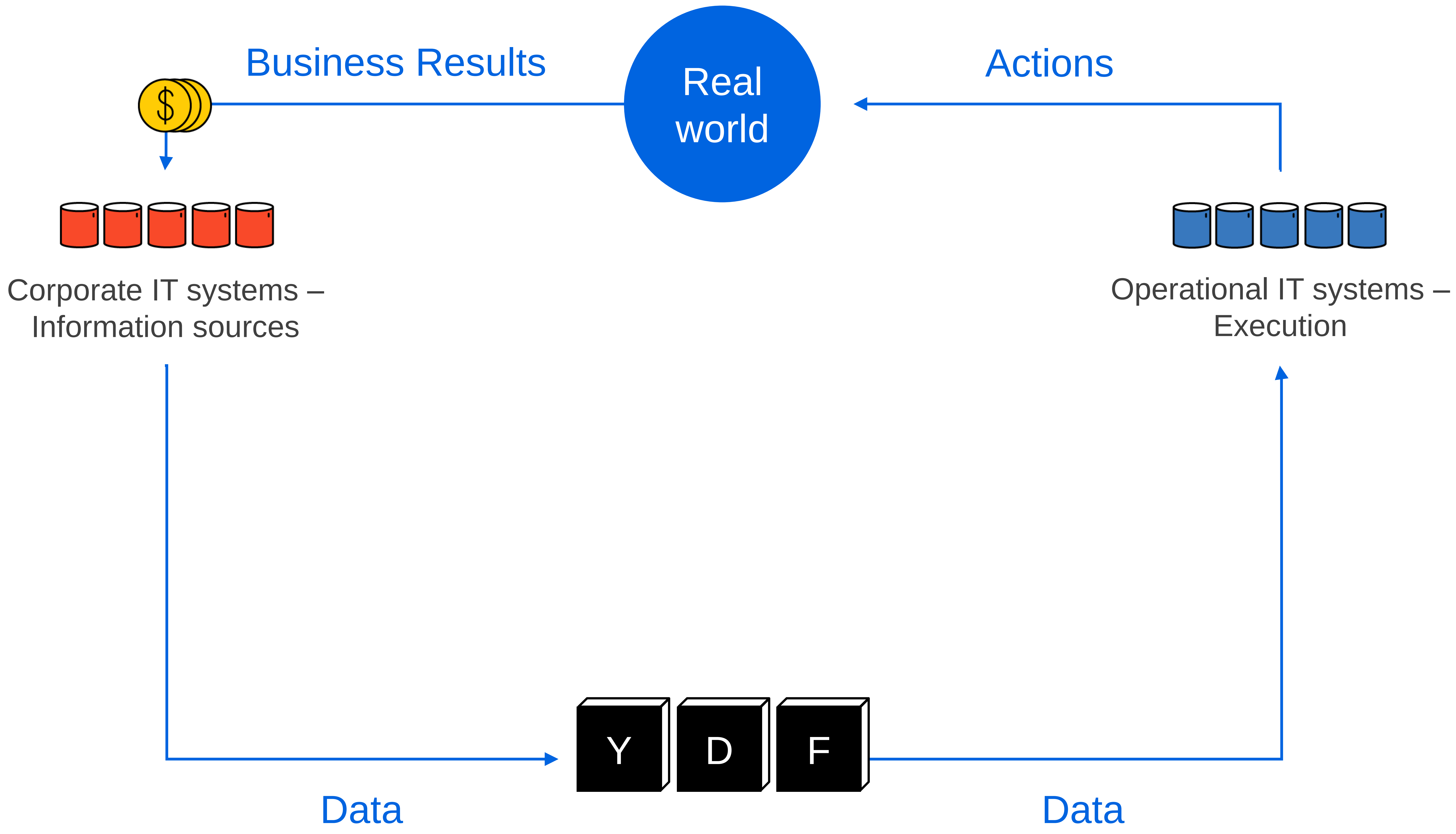
Data: visual in-store display data, CCTV streams, user-generated images, etc.

Solutions: video- and photo stream analytics, automatic planogram compliance, etc.

Typical business process



ML-augumented business process



Yandex
Data Factory

Industries we work with

- Manufacturing
- Retail
- Banking & Finance
- Telecommunications
- Gaming
- Logistics
- Utilities
- Agriculture
- Aerospace
- Healthcare
- Science
  (CERN, neuro science, astrophysics)

Churn prediction*

Customer	Telecom service provider	Mobile operator	WARGAMING.NET LET'S BATTLE Online gaming company
Task	To predict probability of customer churn in the next month(s) to focus retention efforts		
Data used	Logs of > 3,000,000 users for 1 month, data usage only	Logs of services usage of 100,000 users for 3 months	Logs of 100,000 users (game logs, payments logs, etc.) for 1 year
Result	Lift10** = 7.07	up to +11.3% in Lift10 to existing operators model	Lift10 = 6.06

*) *Churn event* – customer leaves a company
**) *Lift* = how many times better the predictive model vs an alternative (random guess).
Lift10 – Lift metric on selection of 10% users that get top churn score by the model
i.e. *Lift10* = (number of predicted churn-ers / real number of churners) for top10% quantile, according to the model prediction

Online recommender system for a fashion ecommerce retailer

Client case

+7%

the average shopping cart value

+35%

the number of products added to customers' shopping carts

Data used

— Logs of visitor behaviour on the website (items viewed, items added to the cart etc.)

— List of goods/products

ML Metrics

— Precision@K (true positive / (true positive + false negative) for top *K* ranked items)

Comparison to competitors

— Bettering the next-best competitor's result by 35% by the number of products added to customers' shopping carts

— The average shopping cart value was 7% higher

— The number of purchases also increased

Optimisation of ferroalloy consumption for a steel production company

Client case

MAGNITOGORSK IRON & STEEL WORKS

5%

average decrease of ferroalloy consumption

>$4.3m

yearly economic effect

Task

To reduce the usage of ferroalloys and other additives in an oxygen-converter plant while maintaining alloy quality standards

Data used

Historical data on more than 200,000 smeltings, incl.:

— Mass of scrap and crude iron

— Steel grades specifications

— Technical parameters of the oxygen-conversion & refining stages

— Results of chemical analyses

— Chemical composition requirements and standards for ferroalloy use

ML Metrics

— Mean Square Error (MSE)

Result

— A service that recommends the optimal consumption of ferroalloys and other materials at a given stage of the production process

— Service integrated with the existing customer software

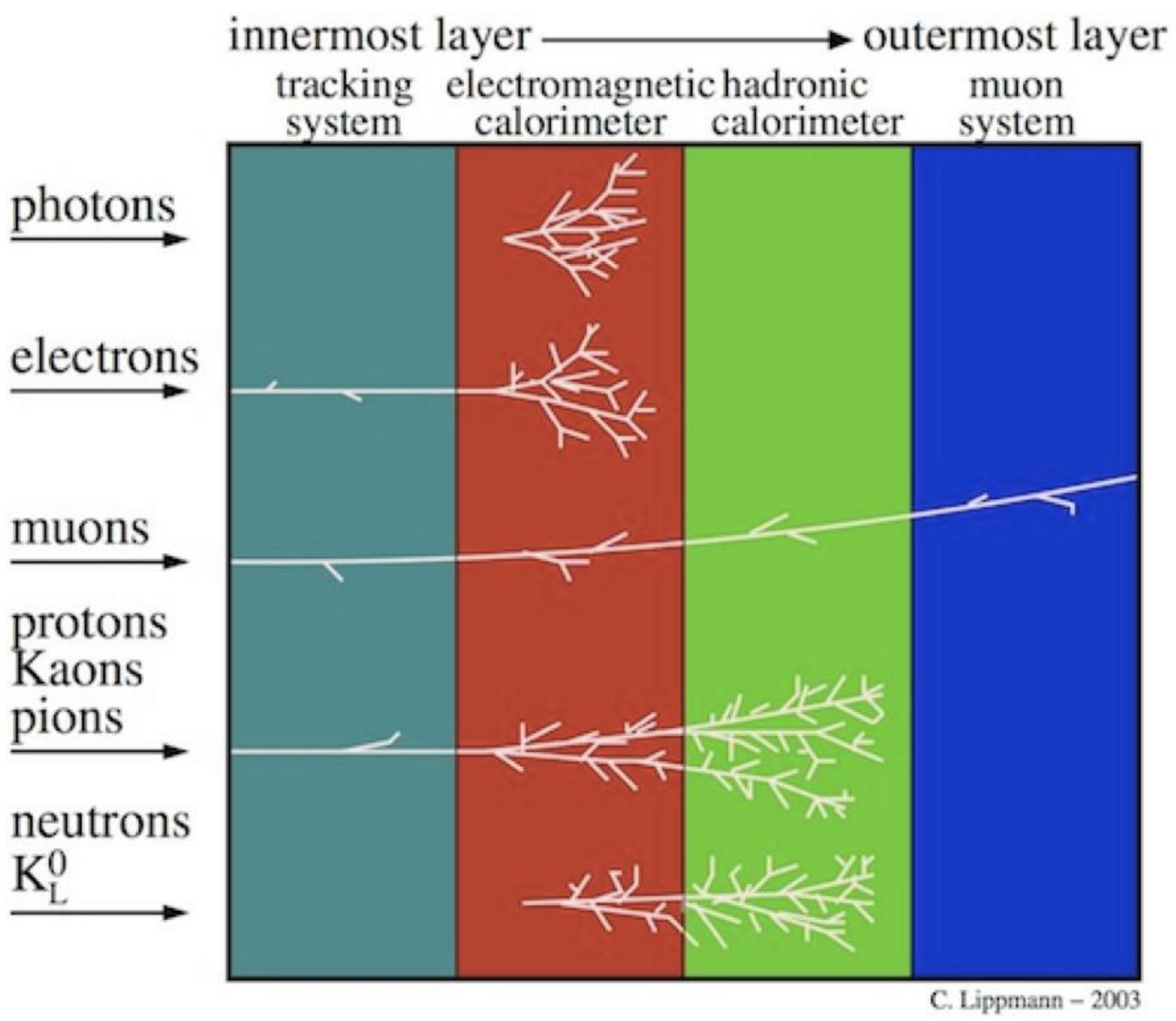
— Reduced consumption of ferroalloys (average of 5%), leading to expected savings of over $4.3m yearly

Other cases

Domain	Solution
Sales and Marketing	— Personalised product or content recommendations — Churn prediction and loyalty management
Operations	— Demand and load prediction
Production and Technical Maintenance	— Production costs optimisation — Quality prediction and predictive maintenance — Diagnostics and monitoring with computer vision
HR	— Personnel screening; prediction of the intention to resign — Candidate scoring for mass recruitment
Other Solutions	— Automated web monitoring — Computer vision based solutions (photo identification, automatic image moderation and classification)

LHCb particle identification

Case



Up to

50%

algorithm error reduction

Task

identify charged particle associated with a track (multiclass classification problem);

particle types: Electron, Muon, Pion, Kaon, Proton and “Ghost”;

combine information from LHCb subdetectors: **CALO**, **RICH**, **Muon** and **Tracker;**

decorrelate (flatten) model output wrt 4 features (P, Pt, eta, nTracks);

Data used

— LHCb Simulated sample

ML Metrics

— ROC AUC one vs all,

— model flatness

Result

— Blended NN model that has error rate half less than baseline for some of the particles;

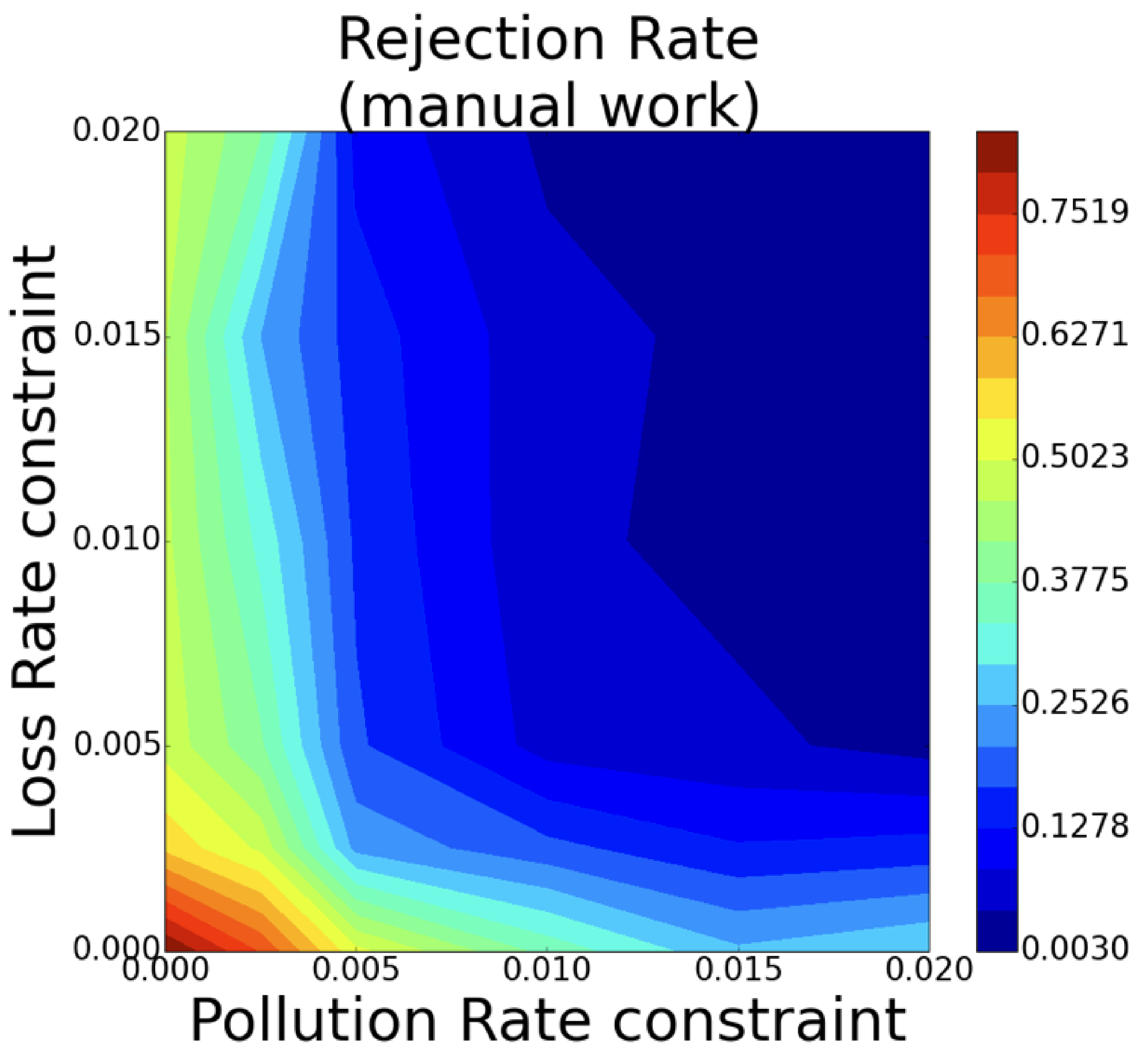
— Blended BDT model with same ROC AUC, but that is flat wrt given features;

— http://bit.ly/2l0yvXc

Yandex Data Factory

CMS data certification / anomaly detection

Case



80%

saving on manual work on data certification tasks

Task

Traditionally, quality of the data at CMS experiment is determined manually. It requires considerable amount of human efforts;

Data used

— CERN open data portal 2010;

— Features: Particle flow jets, Calorimeter Jets, Photons, Muons;

— The dataset was labeled by CMS experts (~3 FTEs).

ML Metrics

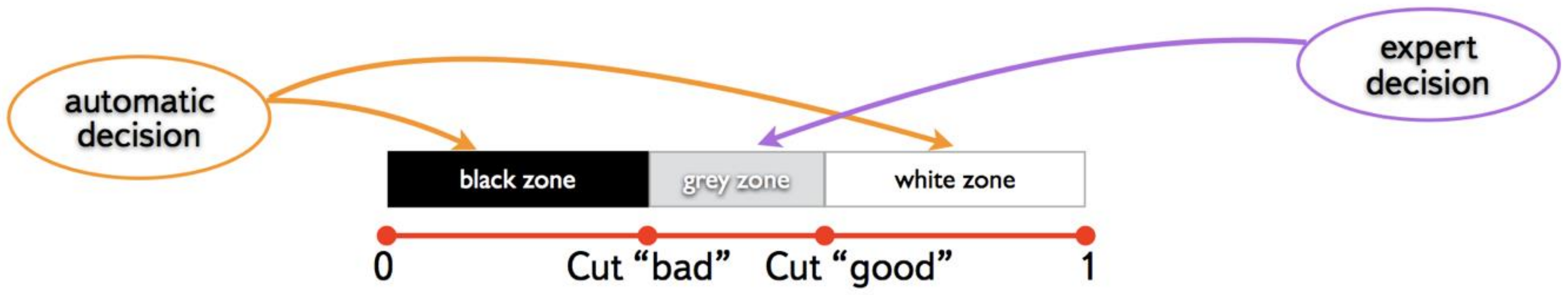
— ROC AUC, precision

Result

— ~80% saving on manual work is feasible for Pollution & Loss rate of 0.5%.

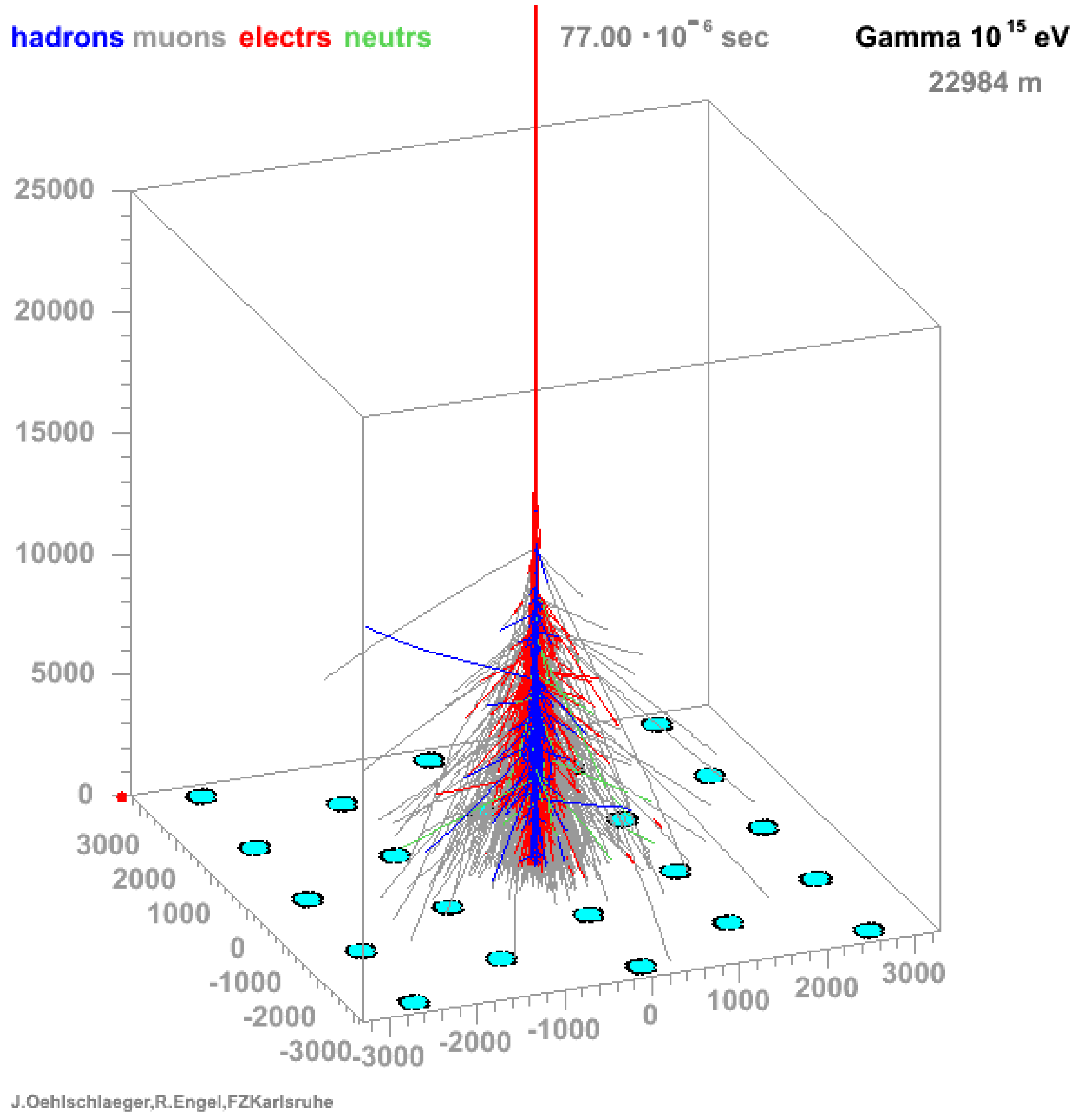
— Next steps: adopt technique for 2016 data & run in production

— http://bit.ly/2l0MLiN



CRAYFIS muon trigger for smartphone

Case



Up to

98%

speedup for running deep neural net model

Task

CRAYFIS experiment proposes usage of private mobile phones for observing Ultra-High Energy Cosmic Rays. Distributed observatory, seeking for particles of energies > 10^18 eV. Design trigger for mobile device that can catch

- an intensive air shower from UHECR (occurs in less than microseconds);
- supports high frame rate (10 Hz)
- trigger on minimally ionizing particles (assuming that such particles leave traces with brightness comparable to the level of intrinsic camera noise).

Data used

- CRAYFIS Simulated sample

ML Metrics

Linear combination of

- weighted cross-entropy;
- computational complexity.

Result

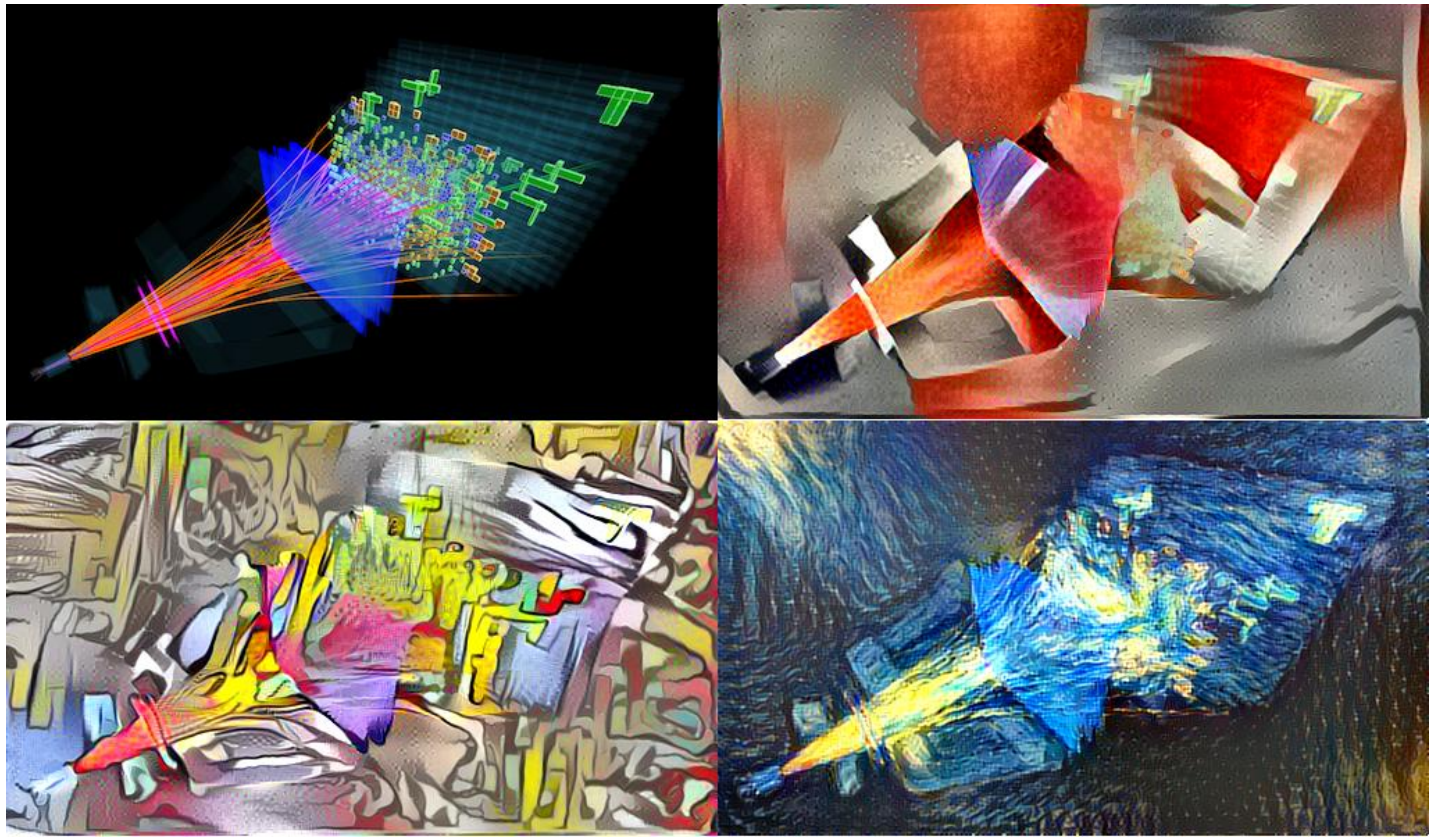
- for just 1.4 times more computational cost than simple cut, gives signal efficiency of 90% and background rejection 60%;
- computational complexity is 0.02 of regular convolutional network;
- http://bit.ly/2nb7gfx

ML task matching

Problem, HEP	Experiment	ML methods	Yandex Product / Task
Particle Identification	LHCb	DNN, classification, advanced Boosting	Search: Image Ranking
MC generation optimization	SHiP	GP, model calibration, non-convex optimisation	Taxi: Multiagent model tuning
Tracking	LHCb, SHiP, COMET	Tracking, Clustering, real-time	Map: Car trajectory identificaiton
Triggers	CRAYFIS	Enhanced Convolutional Neural Nets (CNN)	Navigator: traffic sign identfciatoin
Data modelling	CRAYFIS	Generative Adversarial Nets (GAN)	*To be identified*
Anomaly Detection, data certification	LHCb	Time Series, Binary classification	Anomaly explanation
Detector optimisation	SHiP	Surrogate modelling	Hyper-parameters optimisation for ML model

Conclusion

- If you know Machine Learning, it’s difficult to stay without a job
  - Science: Physics, Medicine, Neurobiology, Molecular Biology, Genomics, …
  - Industry: big companies, startups, consulting;
  - many more stories to be told.
- Key component to success:
  - (Upper) management education;
  - Research-Production pipeline (0-delay deployment);
  - Metric matching from Real World to ML-world.
- Summer School on Machine Learning: MLHEP-2017 Reading, UK, 17-23 July, http://bit.ly/mlhep2017



SCHOOL OF DATA ANALYSIS

Imperial College London

WINTON

Yandex Data Factory

Thank you for attention!

Andrey Ustyuzhanin

anaderiRu @ twitter

anaderi@yandex-team.ru

Backup

Prepaid customers segmentation for a telecom service provider

Client case

Challenge

Client wants to target advertising for his prepaid customers but knows little or nothing about their profiles.

Task

To do segmentation of prepaid (anonymous) telecom customers for further marketing activities based on their data usage history

Data used

— 24 hours history of data usage for 1,900,000 customers

— External web sites catalogue

Result

— Successful segmentation by gender and age group

— Results based on a minimal dataset of 24 hours are comparable in quality with those based on very large historical data

Upsell recommendations for a retail bank

Client case

+13%

increase of additional NPV from the upsell campaign compared to the in-house analytical approach

Task

To provide recommendations for active sales of new services to the existing customers

Data used

— Monthly historical data for 3 million customers over a period of six months (18 million records with about 200 features each)

— Customer’s data included demographics, banking history, history of previous communications, history of banking products purchases by the customer

— Blacklisting rules to determine if certain credit products can be offered to specific customers

Result

13% increase of additional NPV from the upsell campaign compared to the in-house analytical approach

Online recommender system for a large consumer electronics retailer

Client case

Challenge

There is a need to facilitate user's search for the needed item in a large assortment and by this means – to increase conversion rate, revenue and average purchase size. Additionally, this also helps increasing customer loyalty and retention.

+1.5%

revenue per session

+3.2%

conversion rate

Task

To provide product recommendations (including “similar items” block) on the website in order to increase conversion to purchases

Data used

— Website users’ behavior logs (items viewed/added to basket, previous sessions and purchase history)

— Range of goods info (item name, category, price, real-time popularity including purchases made and item page views, promotional offers information) for 58 operation regions of this specific retailer

Result

— Two models designed

〉 Predicting the purchase probability for the specific client

〉 Finding similar items and accessories (for cross-sell)

— Revenue per session +1.5%

— Conversion rate +3.2%

Computer vision based online recommender system for a fashion e-commerce retailer

Client case

Task

- To find similar items in the client's inventory, basing on their visual resemblance
- To provide automatic recommendations of the items to be included in the "Similar Items" block on the web-site, following the business rules on the client's side (category limitations, price limitations, number of items to show, etc.)

Additional tasks

- To find duplicate items in the client's inventory (in order to avoid having identical products listed with different price)
- To find duplicate items among competitors' products (price monitoring)

Project details

- Size of client's collection: 15M images, with 1M images added monthly
- Integration via APIs, service provided in SaaS mode

Result

Recommendations of the relevant similar items on the client's web-site, that allow increasing customers' engagement and conversion rate

TV advertisement budget optimisation for a large consumer brand

Client case

Challenge

Client wants to optimise floating budget for TV advertisement while keeping the target reach values.

2.9%

decrease in floating accommodation budget

$40,000

projected savings per 1 advertising campaign in this client's case

Task

— To forecast what the reach will be for the specified target audience and the budget set using traditional allocation methods

— To build the model that optimises the floating budget:

 〉 By setting how budget should be allocated across different TV advertising channels so that the total budget is reduced to a minimum

 〉 While the target reach values are still met

Data used

— Logs of advertising run times

— TNS data on reach values

— Past advertising campaigns data

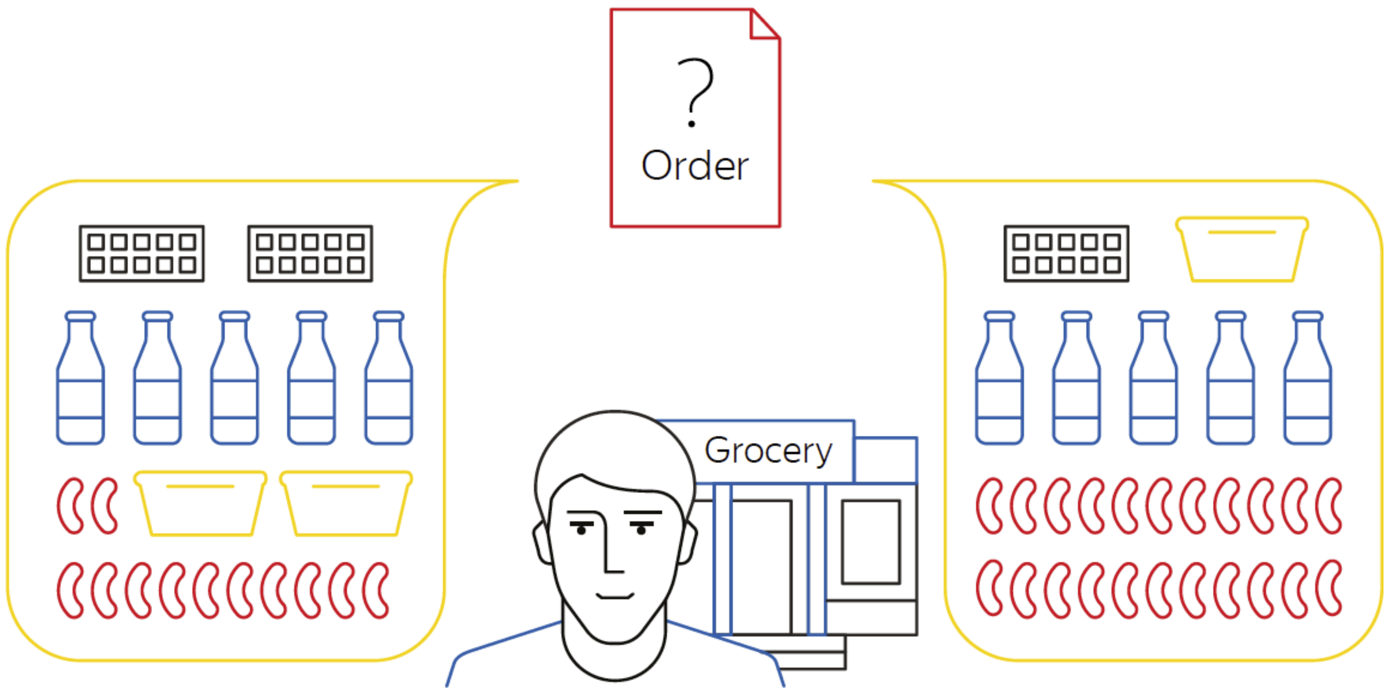
— Schedule of TV advertising spots and advertising rates

Result

— 2.9% decrease in floating accommodation budget

— $40,000 projected savings per 1 advertising campaign in this client's case

Promotion sales forecasting at the retail chain "Pyatyorochka"



About the project

The demand forecasting service for products on retail sales promotions has passed pilot testing at one of Russia's leading grocery chains, Pyatyorochka.

87%

of predictions were accurate within one wholesale package

Task

To accurately forecast demand for products on sales promotion in order to optimise delivery volume (number of packages) for 91 retail stores

— For every product on promotion

— In each store within a retail chain

— A month before the beginning of the promotion itself

Data used

Over 2 years of various sales statistics obtained from receipts, such as store and item ID, region, store type, item category, discount size, item price

Result

— Forecasted the exact number of wholesale packages needed 61% of the time

— 87% of predictions were accurate within one wholesale package

Peak days forecasting for a large retail bank

Client case

Challenge

The client wants to optimise work shifts in order to serve all the customers by an optimal number of employees, thus decreasing overall costs and cutting down the queue times.

+44.3%

increase in accuracy compared to the model previously used

5 out of 6

peak days on average are detected by the model

Task

— To build an individual forecast of peak days for each retail branch with a queue management system implemented

— To predict 5 out of 6 peak days, by month

Data used

— Internal data on client traffic in retail branches (July 2013 - December 2014, daily number of visitors)

— External data on public holidays

— Publicly available information about retail branches (address, working hours, closest metro station, handicapped entrance availability, services provided)

Result

— Accuracy of prediction > 80% (model detects on average 5 out of 6 peak days)

— 44.3% increase in accuracy compared to the model previously used by the bank

— Implementation of the model will allow optimising the number of employees per shift (expected total decrease in man-hours 22%)

ATM cash demand forecasting for Raiffeisenbank

Client case

Challenge

Client wants to increase the precision of cash demand forecasting for ATM network in order to reduce the amount of loaded money while maintaining availability and decreasing the overall costs of replenishment.

15%

expected cost savings

Task

To forecast when and for which amount each ATM within a network should be reloaded

Data used

— Information about ATM (ATM ID, address and type of installation) provided by the bank

— Information about transactions (amount and time) provided by the bank (for 2,000 ATMs)

Result

— 30% decrease in forecast deviation from the actual demand

— Expected savings of 15% of replenishment costs

Passenger traffic prediction for a railway company

Client case

91.3%

forecast accuracy for the passenger traffic for a year for one direction

Task

- Predict passenger traffic (1 year ahead) for one direction
- Predict passenger traffic by classes (business, comfort, economy) and departure time

Data used

- Historical data on passenger traffic
- Data on a schedule and ticket prices
- Railway maps
- Information about trains

Result

Forecast of passenger traffic for a year for one direction, with an accuracy of 91.3%

Optimisation of gas fractionation unit operation for a gas processing company

Client case

Task

Using machine learning technologies, to improve the accuracy of conventional thermodynamic models and identify the factors that decrease the performance of chemical-technological complexes.

1. Identify the typical scenarios and operation modes for the gas fractionation unit
2. For each mode: Identify the factors that affect the performance of unit

Data used

— Data on the hourly operation modes from September 2012 to June 2015

— Data on the composition of the raw materials

— Ambient temperature

Metrics used

— MSE

Result

— A model that allows to identify the factors affecting the unit productivity in different operation modes

Next steps

— Development of the decision support system for the unit operator that provides recommendations on the fractionation unit parameters for maintaining the best performance

— Development of a semi-automatic/automatic system to control gas fractionation unit parameters

Prediction of engineers' resignation for a large R&D centre

Client case

26 out of 50

predicted resignations actually happened

$0.5m

in expected savings for top-50 of each 1,000 engineer positions analysed

Task

— To save costs for research and employment of high level engineers by predicting their intention to resign

— To compare the prediction results with the real resignation statistics

— To evaluate the "weight" of different factors in this prediction

Data used

— Resignation statistics within the group of 1255 engineers for the period from 2008 till 2012

— Employees data from HR systems (gender, age, education, trainings passed, vacations frequency, etc.)

Result

— Model providing a forecast of the employees who are going to voluntary resign in the next time period

— 26 out of 50 employees indicated as those with the highest probability to resign actually left the company

— Potential savings up to $0.5m for Top-50 of each 1,000 engineer positions analysed

— The strongest factors identified

Automatic image moderation and duplicate search for an online social service

Client case

88.38%

total accuracy of classification into 7 custom categories

Task

— To minimise the costs of moderating the content that is uploaded to the website and accelerate this process through its automation

— To classify images to 6 categories: ‘children’, ‘adult content’, ‘women’, ‘men’, ‘many’, ‘other’

— To find duplicates of images both in client’s data and online in order to recognise fake photos

Data used

— Training set of images with known characteristics

Result

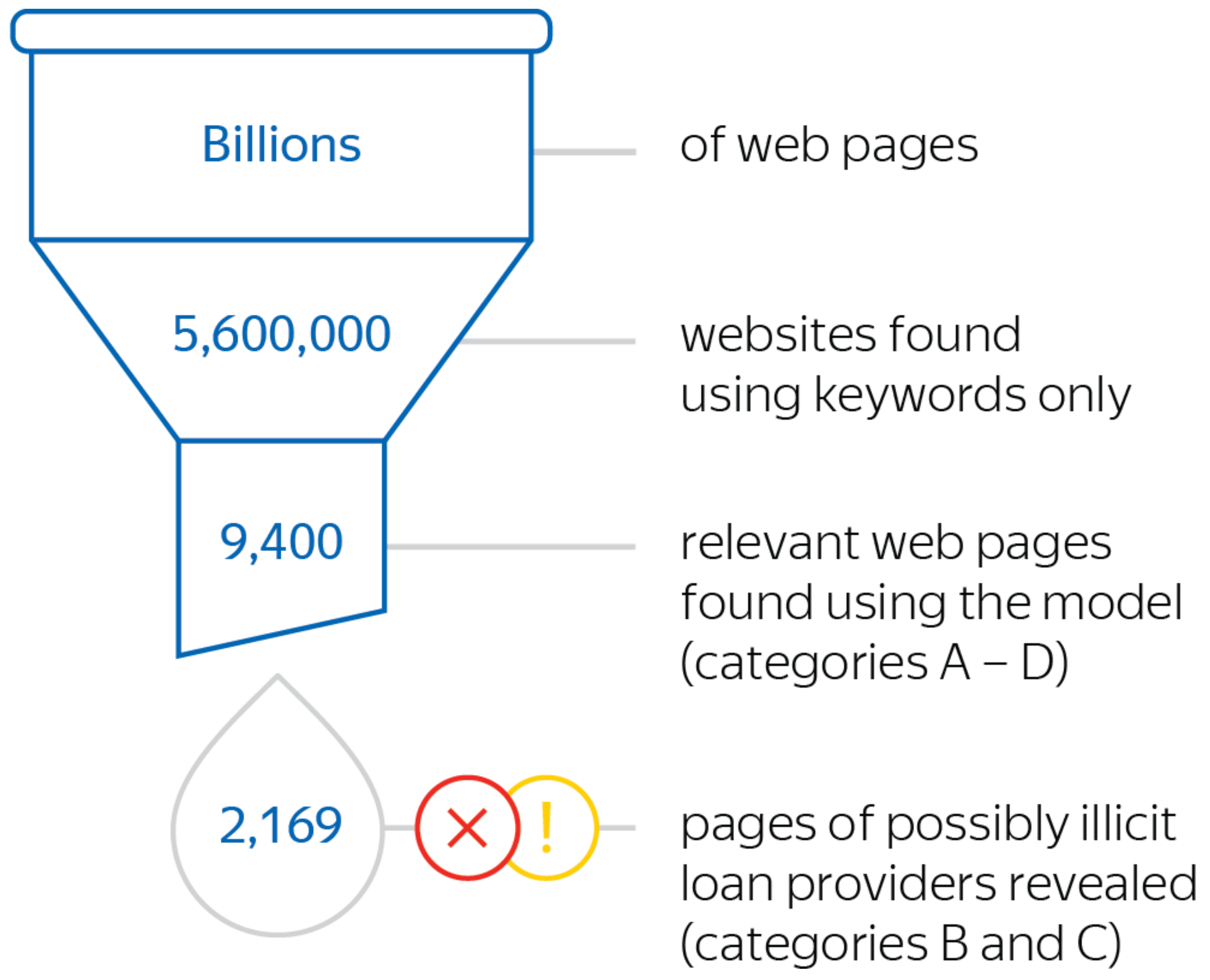
Service helped to cut the client’s image moderation costs through:

— Classifying images into 7 categories with total accuracy of 88.38%

— Exposing images with large number of duplicates online and within client’s data

Automated web monitoring for the Bank of Russia

Client case



9 out of 10

pages that most likely belong to illegal loan providers get found

Task

- To find web pages of organisations providing cash loans
- Classify these resources by their likely status (whether they comply with regulatory requirements or not)

Data used

- Yandex search index
- State register of microfinance organisations and legal information on organisations (taxpayer identification number, primary state registration number)
- Website examples marked up by Bank of Russia specialists

Result

A service that:

- Provides a regularly updated database of websites and social media pages that almost certainly belong to loan providers
- Classifies them by their status with
  - 90% recall rate (9/10 pages that most likely belong to illegal providers are found)
  - 98% precision rate (only 2/100 pages marked as “entered in a state register of microfinance organisations“ and “not related” are marked incorrectly)
  - 71% of relevant web pages end up in the correct category

Topic-specific search service on antibiotic resistance

Client case

AstraZeneca

94.8%

recall rate (more than 94 pages out of 100 relevant pages are found)

Task

— Find all scientific documents on regional antibiotic resistance with at least 20% precision and rank them according to their relevance, in order to create an information service for drug-store chains and distributors

— Monitor the appearance of new documents

Data used

— Training set of documents assessed by AstraZeneca experts

— List of keywords and requirements (e.g. only original papers, only those with quantitative results)

Result

— A search service optimised for a specific pharmaceutical topic

— Search quality of 57.9% precision and 94.8% recall (on the first 1000 of documents)

Traffic and road accidents prediction for a road management agency

Client case

FEDERAL ROAD AGENCY
ROSAVTODOR

7x

better accidents prediction compared to average frequency analysis

Task

Predict traffic (1 hour ahead) and road accidents (4 hours ahead) for road management efforts optimisation

Data used

— Historical and actual data

— Weather archive

— Technical parameters of roads

Result

— Traffic prediction of high precision

— Prediction of road accidents 7x better than average frequency analysis

— Model allows improving the road patrol's allocation and reaction times

Infrastructure optimisation for CERN

CERN openlab

As an associate member of CERN openlab, Yandex collaborates with CERN on a number of projects of various scale. Working with CERN, Yandex Data Factory helps to reduce costs to the most famous and promising scientific venture of our era.

Data storage optimisation by 40%

Saving up to 4PB of storage a year, which costs ~$4M

– A similar solution can be performed for the companies that need to store large datasets: prediction of dataset “popularity” and automated decision on type of storage.

Online event filtering efficiency optimisation by up to 60%

Physicists may get same results up to ~1.5 times faster

Detection of anomalies in data allowing to decrease man-hours by 75%

An automatic anomaly-detection service that can save approx. 4,000 man-hours yearly on the CMS experiment.

Data storage optimisation

Client case

CERN openlab

Challenge

The Large Hadron Collider detectors take captures of every notable particle interaction, which piles up to over 5 PB of data a year. But the market price for 1 PB of data storage per year runs as high as 1 million dollars.

$4m

maximum yearly projected savings on data storage

Task

To cut storage costs and to determine which files should be stored on which kind of medium, to improve the effectiveness of data access

Data used

— Historical data on the access history of every file generated by LHCb and the collision simulators (each file catalogued by several features, like file size, number of existing file copies, access frequency, longest duration for which the files hadn’t been accessed, file origin, etc.)

Result

— Data storage optimisation by 40%

— A model that allows saving up to 4 petabytes (more than 4 million gigabytes) of storage a year – the standard rate for storage is $4m, annually

— The model has been deployed at the beginning of the Collider’s Run-II in the Summer of 2015

Diagnostic platform in oncology

Client case

AstraZeneca

About the project

Together with AstraZeneca and the Russian Society of Clinical Oncology, Yandex Data Factory launched RAY, diagnostic platform in oncology. The goal is to contribute to better cancer diagnostics methods, as well as to the identification of pre-dispositions to cancers in the Russian Federation.

Task

To create a platform that simplifies access to reference information and helps molecular biologists and clinical geneticists to interpret test results, thus:

— Generates a report on mutations found in a patient's genome

— Delivers information on their potential effects and available cancer treatment options

Data used

— Data on mutations from curated databases, including their clinical significance and relations to diseases

— Publicly available whole-genome sequencing data

— Data on newly investigated mutations added regularly to improve the report comprehensiveness

Result

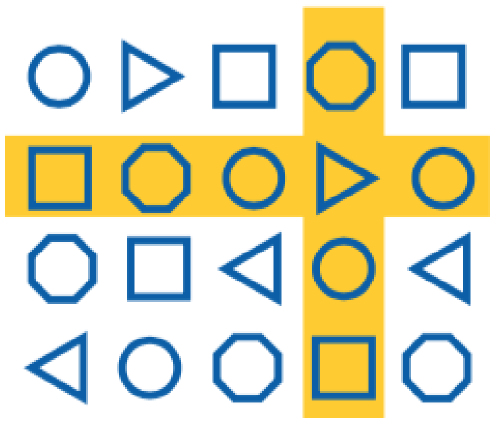
— The testing stage started in December 2015

— The number of tests within the OVATAR study is expected to reach 3,000-5,000 annually at the initial stage and 10,000+ later

06

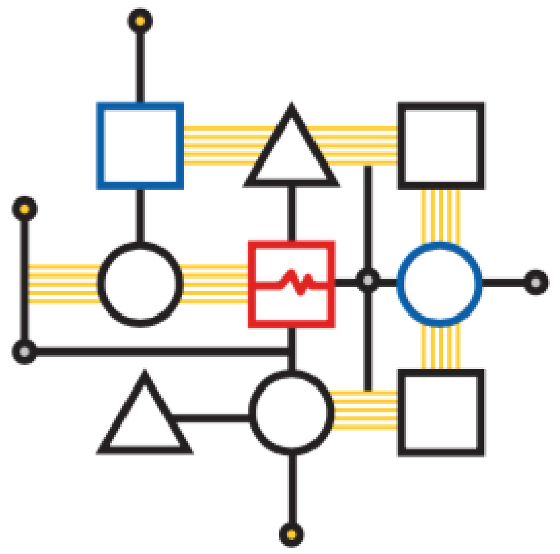
How we work with clients

Project implementation phases



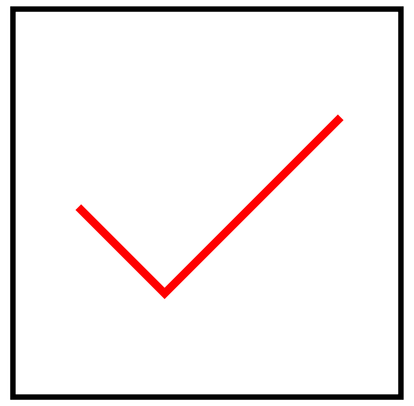
1 Preliminary phase

- Data assessment
- Defining service requirements, metrics and success criteria
- Data transfer (incl. depersonalisation and obfuscation of sensitive data)
- Agreement on experiment procedure



2 Pilot phase and performance evaluation

- Model training
- Model testing (running an experiment)
- Checking success criteria, model performance evaluation



3 Production use

- Regular data transfer or deployment of the model on client's premises
- Regular model quality checks through A/B testing
- Model quality maintenance, including updating the model as new data is received

Yandex Data Factory solutions benefits

As a deliverable, we provide the client with a trained and tested service (software), focused on solving one exact business goal.

√ **Flexible Deployment Schema**
The service can be used as a SaaS, or deployed directly on client's premises.

√ **Quality Improvement**
You'll see our service performance gets better and better with time. The model is constantly retrained based on new data, which in turn improves its quality.

√ **Result Guaranteed**
We guarantee results: you only pay if the value of the predefined quality metric stated in your SLA is met.

√ **Easy Integration via APIs**
YDF's services are easily integrated with standard data sources such as Hadoop or SQL-based data storages, as well as internal systems such as ERP, MES etc.

Data privacy and security

On-premises deployment option

Possible options:

SaaS mode
In this case, both model building and real-time data processing happen in the YDF cloud (best price/effectiveness ratio for most cases)

Hybrid model
Resource-intensive machine learning happens in the YDF cloud, but the resulting model is deployed on the client's premises, reusing existing infrastructure.

Secure storage and data transfer

We provide data replication, data storage in different geographical regions, data backup, end-to-end encryption for data transfer, access control and maximum isolation of data belonging to different clients.

Several data centres assure uninterrupted performance of YDF's services.

Sensitive data handling

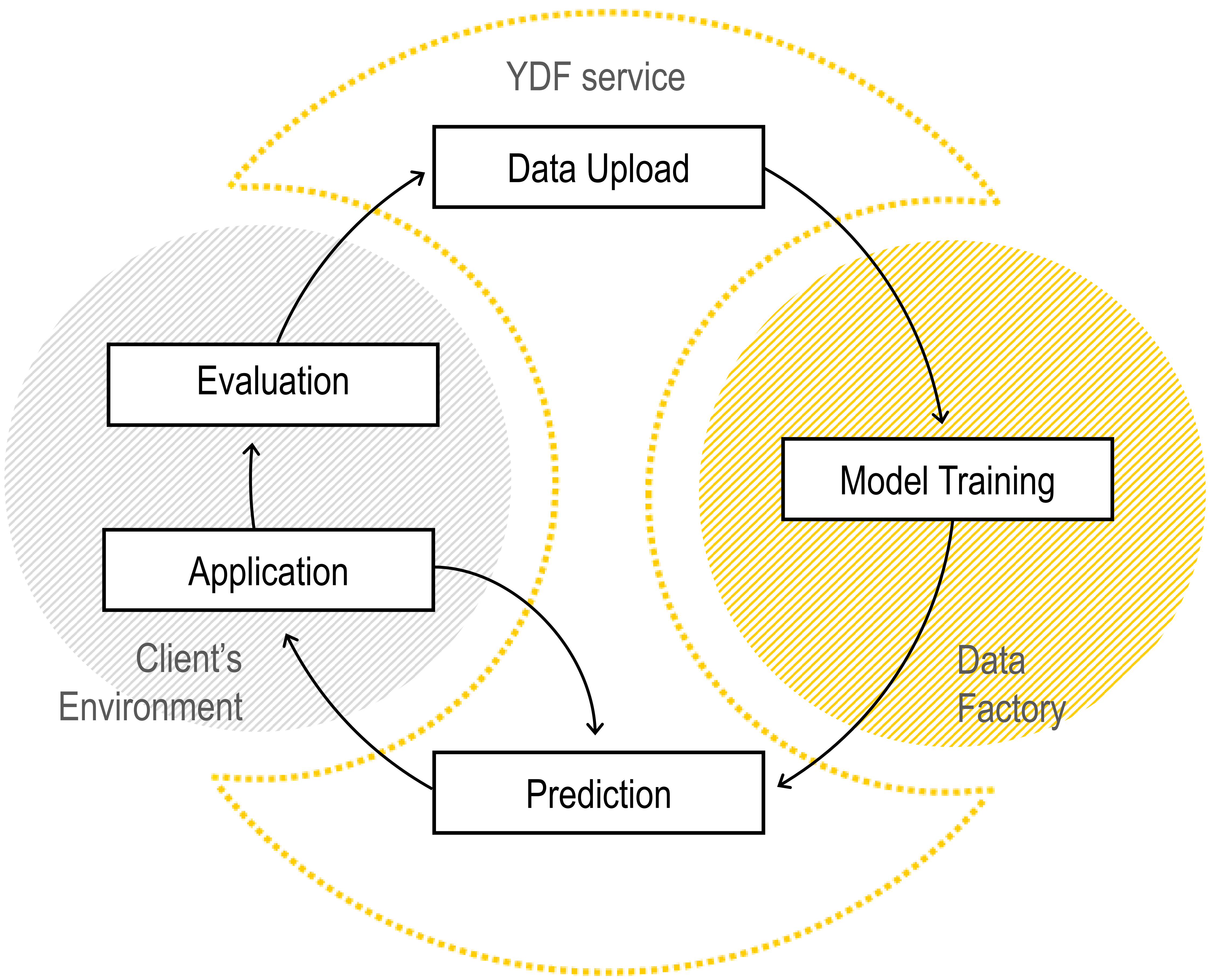
Confidential and personal data go through the process of depersonalisation and obfuscation, which allows decreasing their sensitivity for data transfer.

For regionally sensitive data, we have several data centres in the CIS, EU and Turkey.

Deployed model cycle

Steps:

1. **Data Upload**
   Data uploading and transforming
2. **Model Training**
   Model training based on uploaded data
3. **Prediction**
   Creating predictions and prescriptions based on client's request
4. **Application**
   Prediction and prescription applying
5. **Evaluation**
   Measurement of result, comparison with control sample



Yandex
Data Factory

Tel: +31 (0) 20 206 69 70
+31 (0) 20 206 69 71
Fax: +31 (0) 20 446 63 72
Mail: ydf-customer@yandex-team.com

Yandex Europe B.V.
WTC Schiphol Airport
Schiphol Boulevard 165
1118 BG Schiphol
Netherlands

yandexdatafactory.com