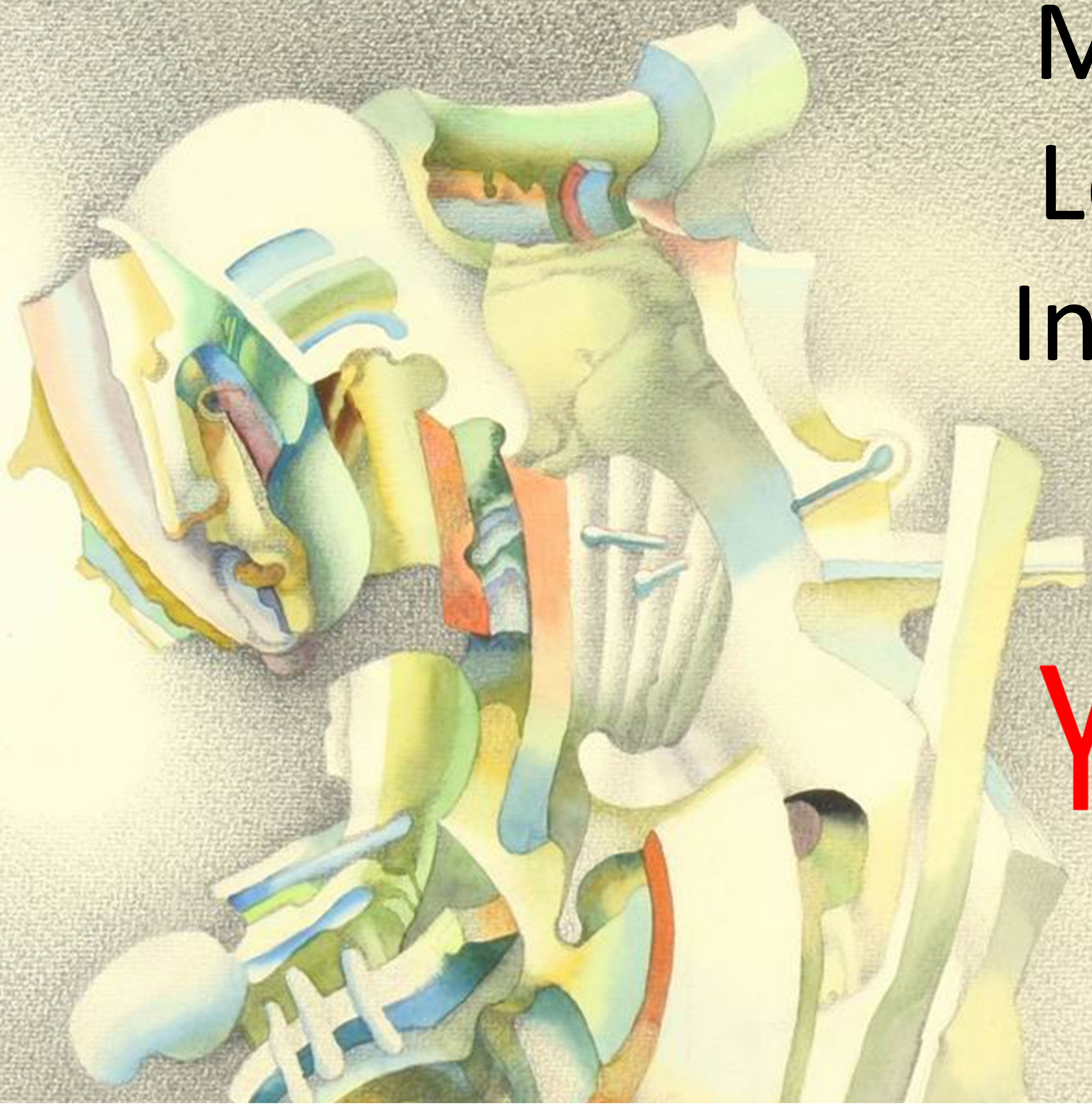
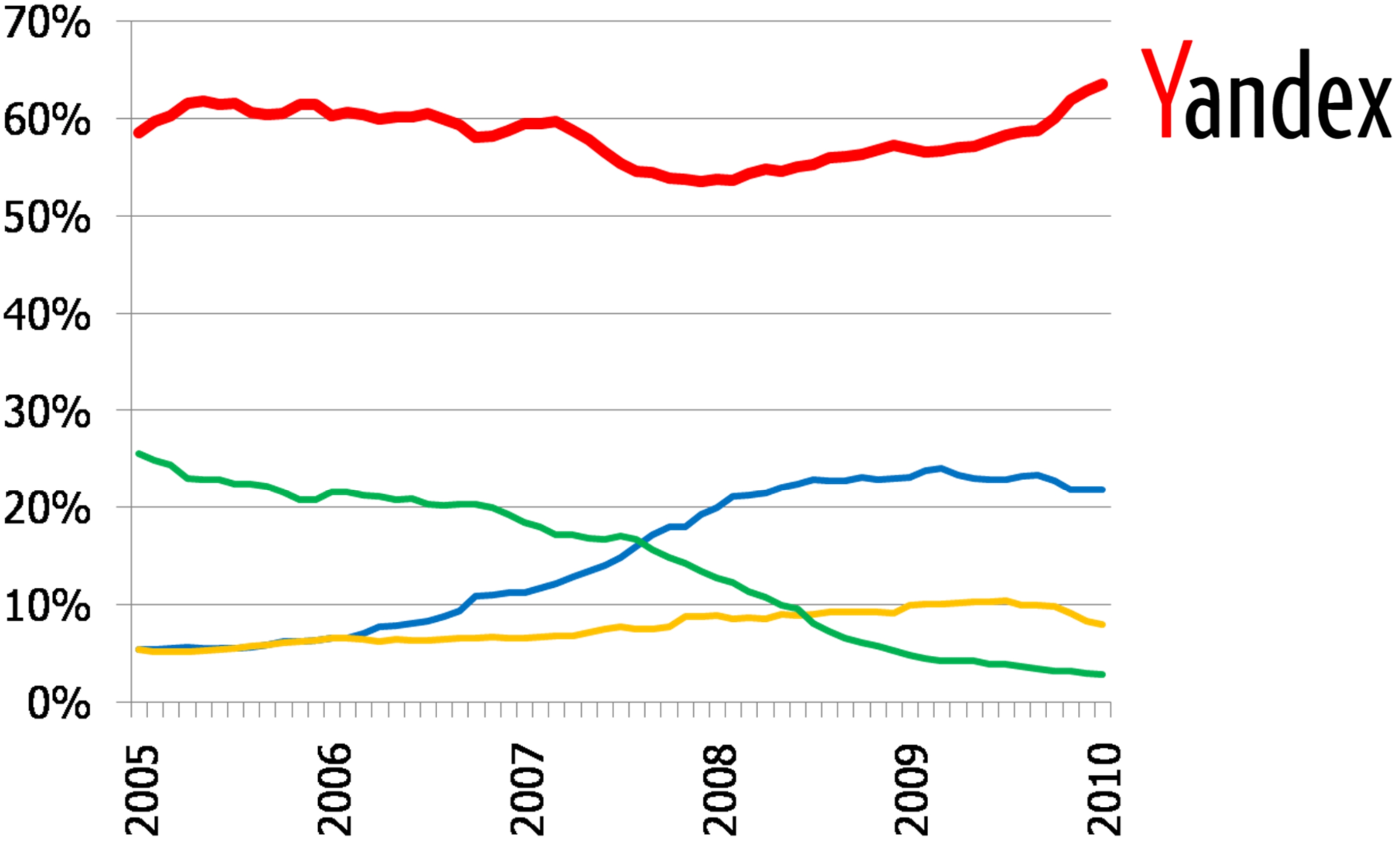


Machine Learning In Search Quality At Yandex

Russian Search Market



Source: **LiveInternet.ru**, 2005-2009

Yandex

A Yandex Overview

1997

Yandex.ru was launched

№7

Search engine in the world * (# of queries)

150 mln

Search Queries a Day

Offices

>Moscow

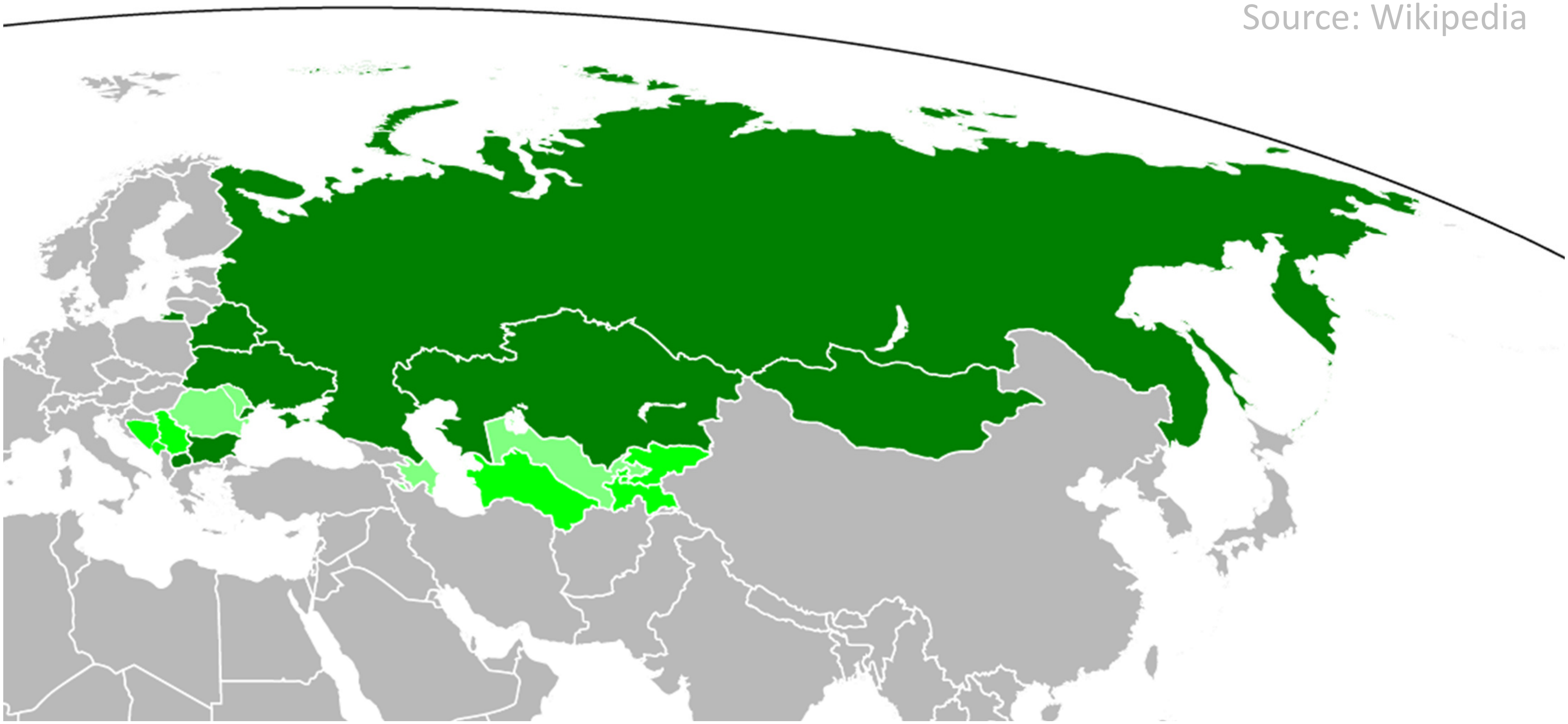
>4 Offices in Russia

>3 Offices in Ukraine

>**Palo Alto (CA, USA)**

* Source: **Comscore** 2009

Variety of Markets



Source: Wikipedia

15 countries with cyrillic alphabet

77 regions in Russia

Variety of Markets

> Different culture, standard of living, average income
for example, Moscow, Magadan, Saratov

> Large semi-autonomous ethnic groups
Tatar, Chechen, Bashkir

> Neighboring bilingual markets
Ukraine, Kazakhstan, Belarus

Geo-specific queries

Relevant result sets vary

across all regions and countries

[**wedding cake**]

[**gas prices**]

[**mobile phone repair**]

[**пицца**] Guess what it is?

pFound

A Probabilistic Measure of User Satisfaction

Probability of User Satisfaction

Optimization goal at Yandex since 2007

> *pFound* – **P**robability of an answer to be **FOUND**

> *pBreak* – **P**robability of abandonment at each position (**BREAK**)

> *pRel* – **P**robability of user satisfaction at a given position (**REL**evance)

$$pFound = \sum_{r=1}^{n} (1 - pBreak)^{r-1} pRel_r \prod_{i=1}^{r-1} (1 - pRel_i)$$

Similar to ERR, **Chapelle**, 2009, Expected Reciprocal Rank for Graded Relevance

Geo-Specific Ranking

An initial approach

query → query + user's region

Ranking feature e.g.: "user's region and document region coincide"

An initial approach

query → query + user’s region

Problems	Hard to perfect single ranking	Cache hit degradation
	> Very poor local sites in some regions > Some features (e.g. links) missing > Countries (high-level regions) are very specific	> Twice as much queries

Alternatives In Regionalization

Separated local indices	VS	Unified index with geo-coded pages
One query	VS	Two queries: original and modified (e.g. +city name)
Query-based local intent detection	VS	Results-based local intent detection
Single ranking function	VS	Co-ranking and re-ranking of local results
Train one formula on a single pool	VS	Train many formulas on local pools

Why use MLR?

Machine Learning as a Conveyer

> Each region requires its ranking

Very labor-intensive to construct

> Lots of ranking features are deployed monthly

MLR allows faster updates

> Some query classes require specific ranking

Music, shopping, etc

MatrixNet

A Learning to Rank Method

Yandex

MatrixNet

A Learning Method

> boosting based on decision trees
We use oblivious trees (i.e. “matrices”)

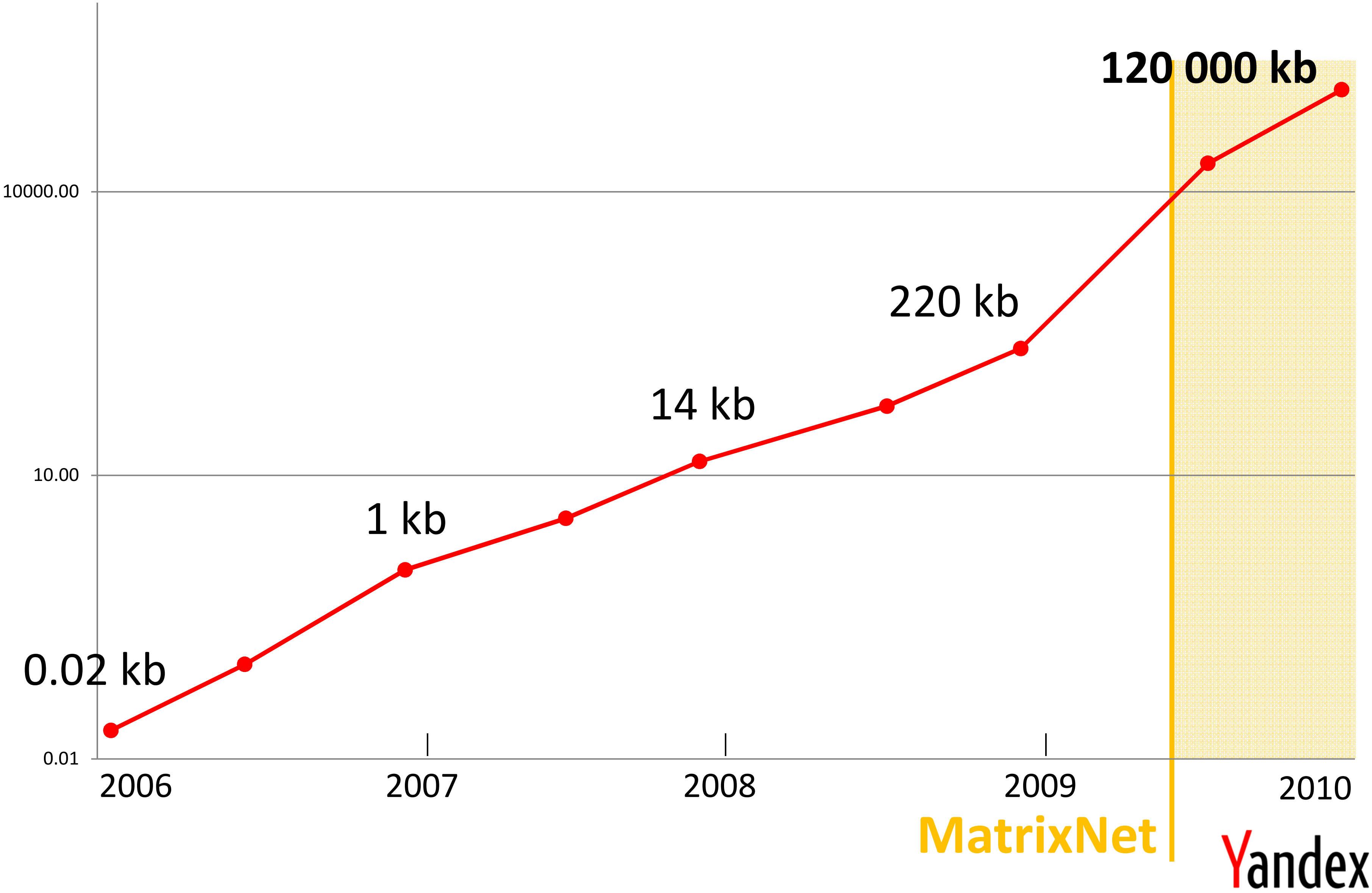
> optimize for pFound

> solve regression tasks

> train classifiers

Based on **Friedman**’s Gradient Boosting Machine, Friedman 2000

MLR: complication of ranking formulas



MLR: complication of ranking formulas

A Sequence of More and More Complex Rankers

> pruning with the Static Rank (static features)

> use of simple dynamic features (such as BM25 etc)

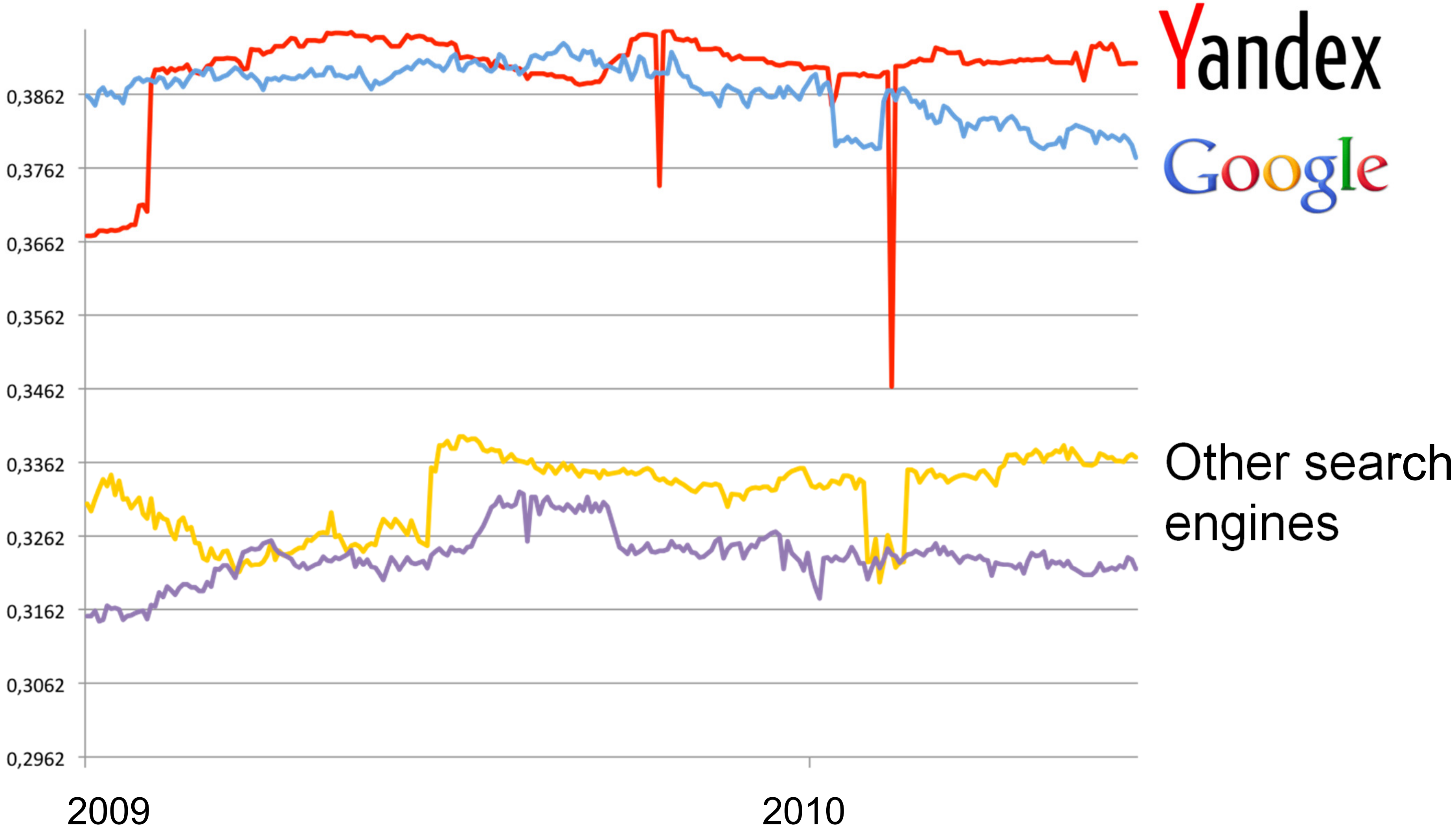
> complex formula that uses all the features available

> potentially up to a million of matrices/trees for the very top documents

See also **Cambazoglu**, 2010, Early Exit Optimizations for Additive Machine Learned Ranking Systems

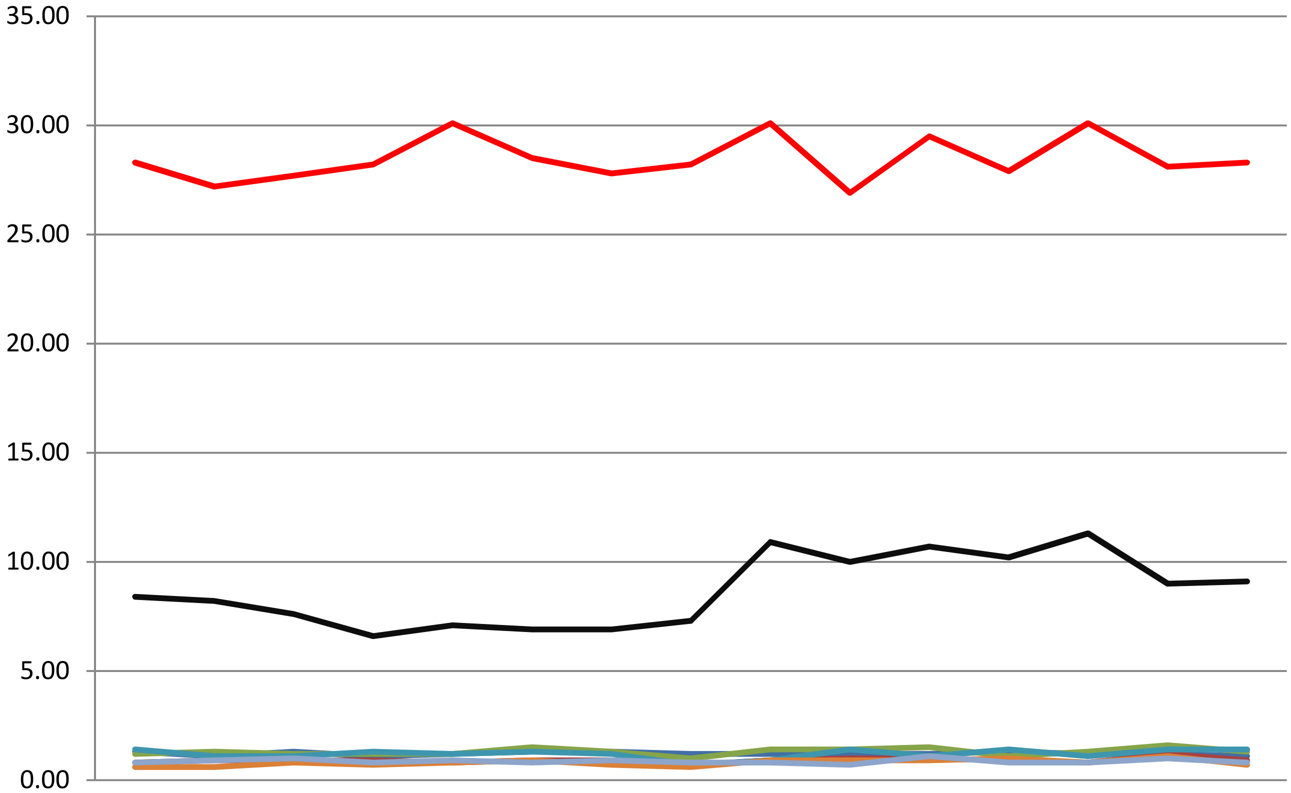
Yandex

Geo-Dependent Queries: pfound



Geo-Dependent Queries

Number of Local Results (%)



Yandex

Player #2

Other search engines

Source:**AnalyzeThis**.ru, 2005-2009

Lessons

MLR is the only key to **regional search**: it provides us the possibility of tuning many geo-specific models at the same time

Challenges

> **Complexity** of the models is increasing rapidly
Don’t fit into memory!

> MLR in its current setting does not fit well to **time-specific queries**
Features of the fresh content are very sparse and temporal

> **Opacity of results of the MLR**
The back side of Machine Learning

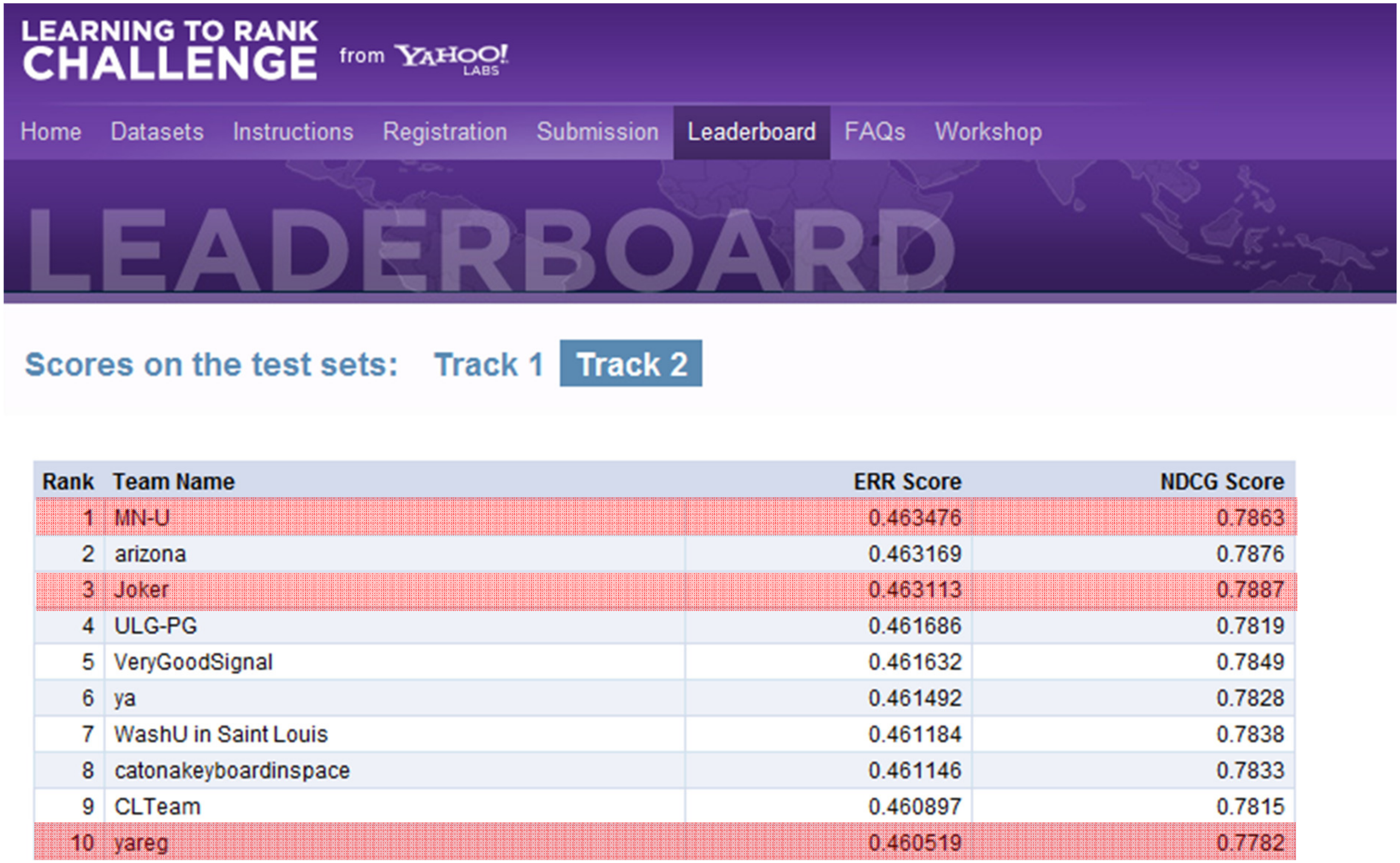
> Number of features grows faster than the number of judgments
Hard to train ranking

> Learning from clicks and user behavior is hard
Tens of Gb of data per a day!

Yandex and IR

Participation and Support

Yandex MLR at IR Contests



Rank	Team Name	ERR Score	NDCG Score
1	MN-U	0.463476	0.7863
2	arizona	0.463169	0.7876
3	Joker	0.463113	0.7887
4	ULG-PG	0.461686	0.7819
5	VeryGoodSignal	0.461632	0.7849
6	ya	0.461492	0.7828
7	WashU in Saint Louis	0.461184	0.7838
8	catonakeyboardinspace	0.461146	0.7833
9	CLTeam	0.460897	0.7815
10	yareg	0.460519	0.7782

№1

MatrixNet at Yahoo Challenge: #1, 3, 10 (Track 2), also BagBoo, AG

Support of Russian IR

Schools and Conferences

>**RuSSIR**, since 2007, – Russian Summer School for Information Retrieval

>**ROMIP**, since 2003, – Russian Information Retrieval Evaluation Workshop: 7 teams, 2 tracks in 2003; 20 teams, 11 tracks in 2009

>**Yandex School of Data Analysis**, since 2007 – 2 years master program

Grants and Online Contests

>**IMAT (Internet Mathematics) 2005, 2007** – Yandex Research Grants; 9 data sets

>**IMAT 2009** – Learning To Rank (in a modern setup: test set is 10000 queries and ~100000 judgments, no raw data)

>**IMAT 2010** – Road Traffic Prediction

http://company.yandex.ru/academic/grant/datasets_description.xml
http://imat2009.yandex.ru/datasets
http://www.romip.ru

Yandex

Yandex

We are hiring!