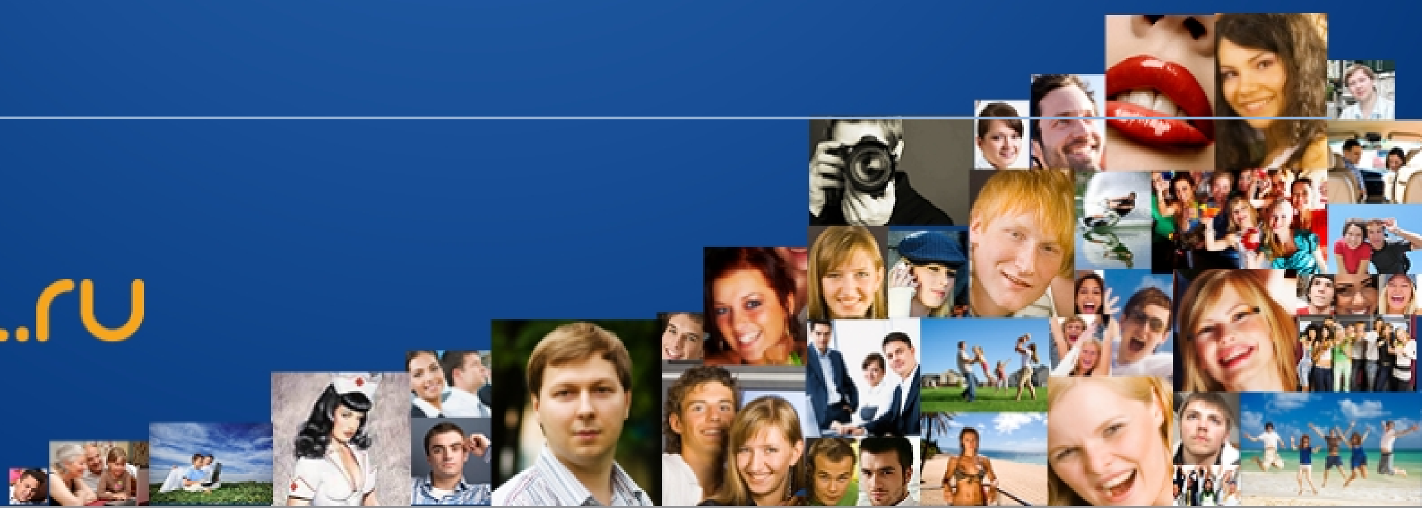
поиск@mail.ru

Introduction to

Active Learning

Alexey Voropaev

go.mail.ru



Mail.Ru have own Search Engine

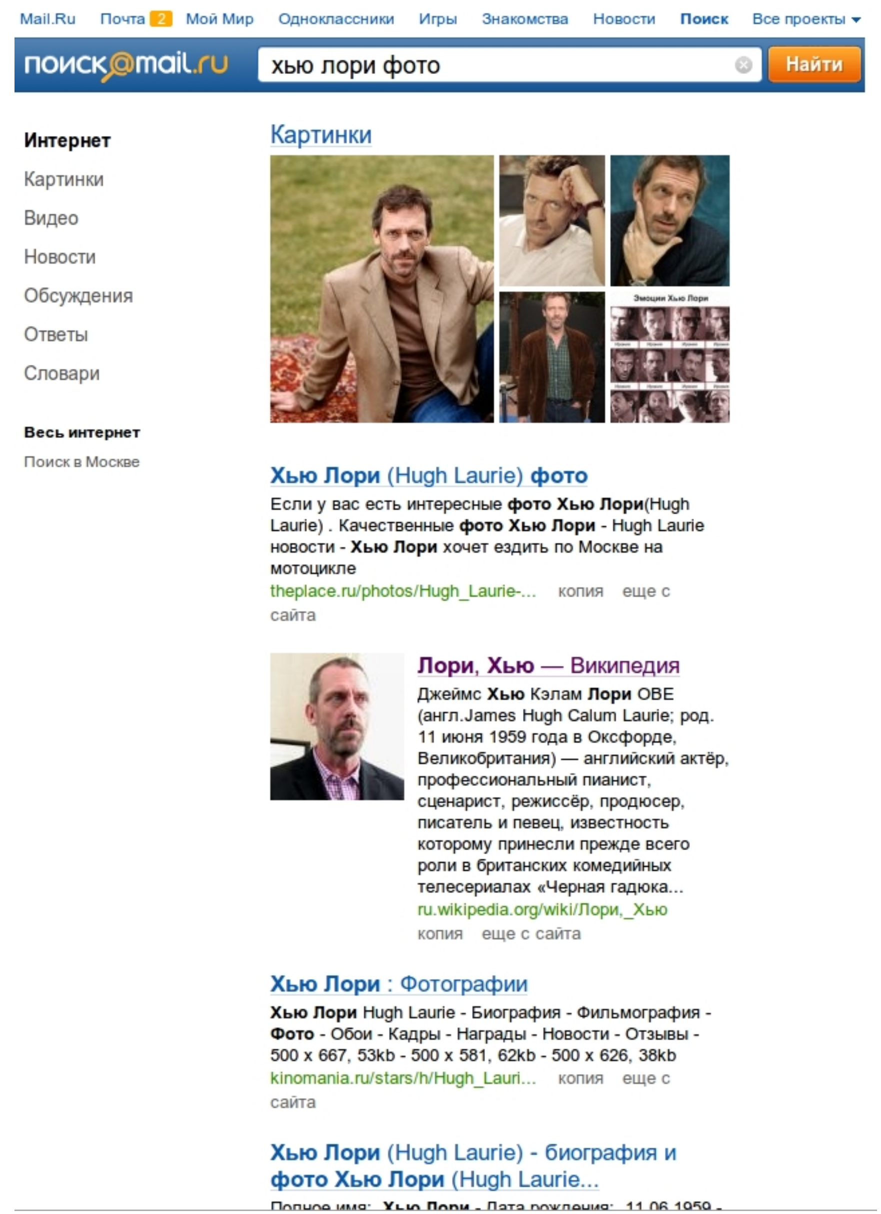
go.mail.ru

– It is not «patched Google»

– **Completely out product**

– 8.3% of market

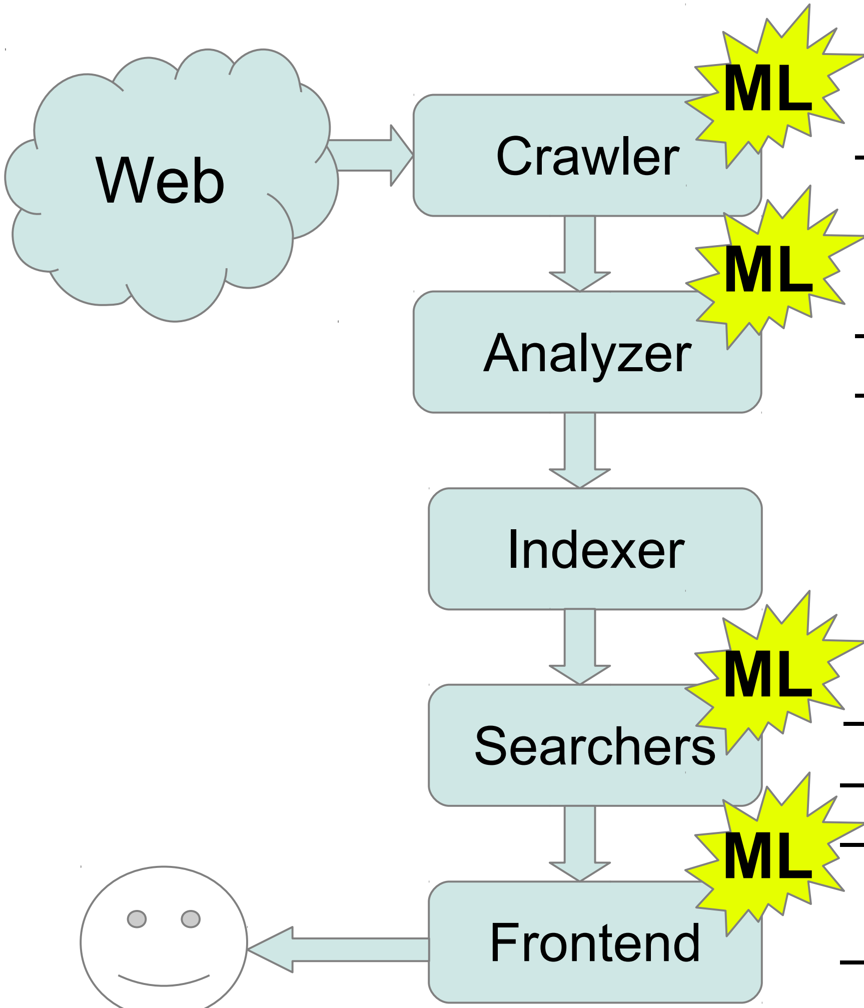
- The same engine for:
    - web
    - image
    - video
    - news
    - real time
    - e-mail search
    - etc.



We love Machine Learning :)

go.mail.ru

Simplified Search Engine Architecture:



– StaticRank: crawling priority

– Antispam
– Antiporn
– Object extraction
– User behaviour

– Ranking
– Lingustic
– Query processing

– Antirobot
– Decision when to show vertical search

«Machine learning is the hot new thing»

John Hennessy, President, Stanford

There is unknown function $f(\vec{x}) = y$

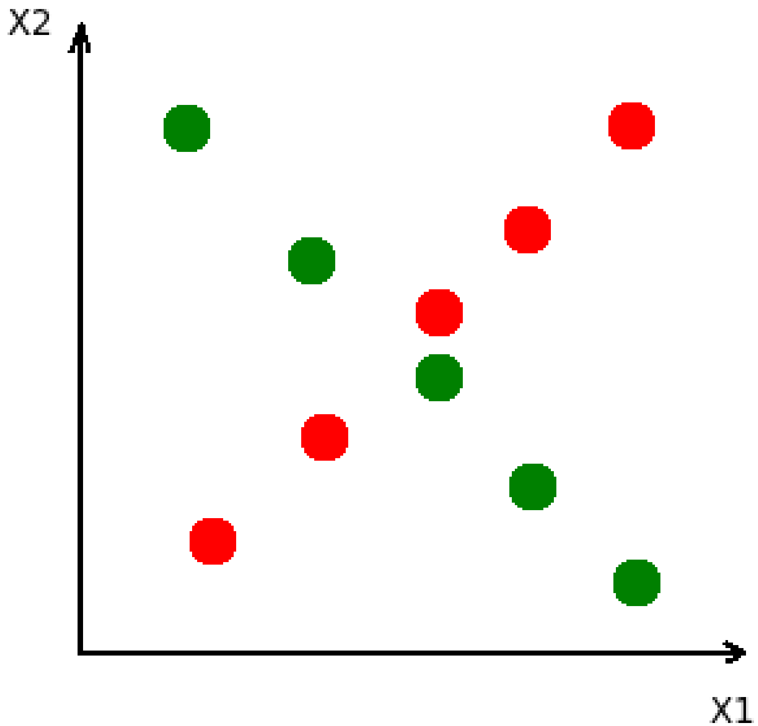
Given:

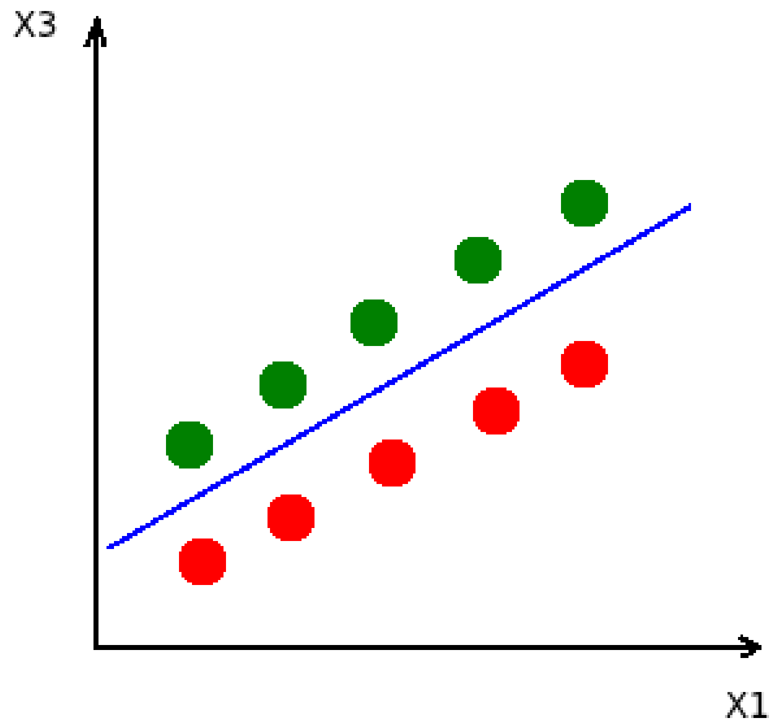
Set of samples $T = \{\vec{x}_1, y_1\} \ldots \{\vec{x}_n, y_n\}$

Goal:

Build approximation f' of f using T

Take good features





A — impossible

B — possible

Lenear model

Proper model selection

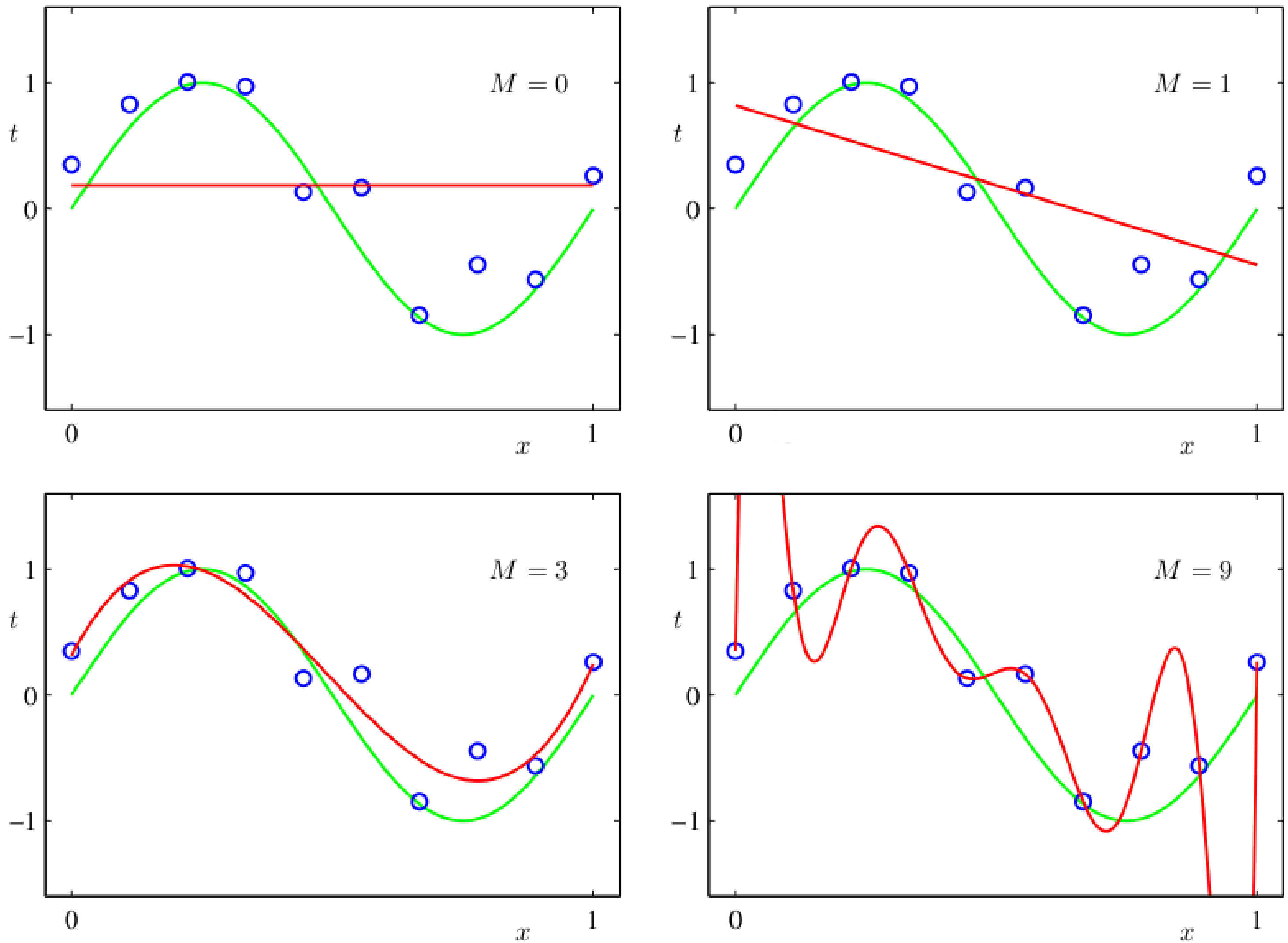
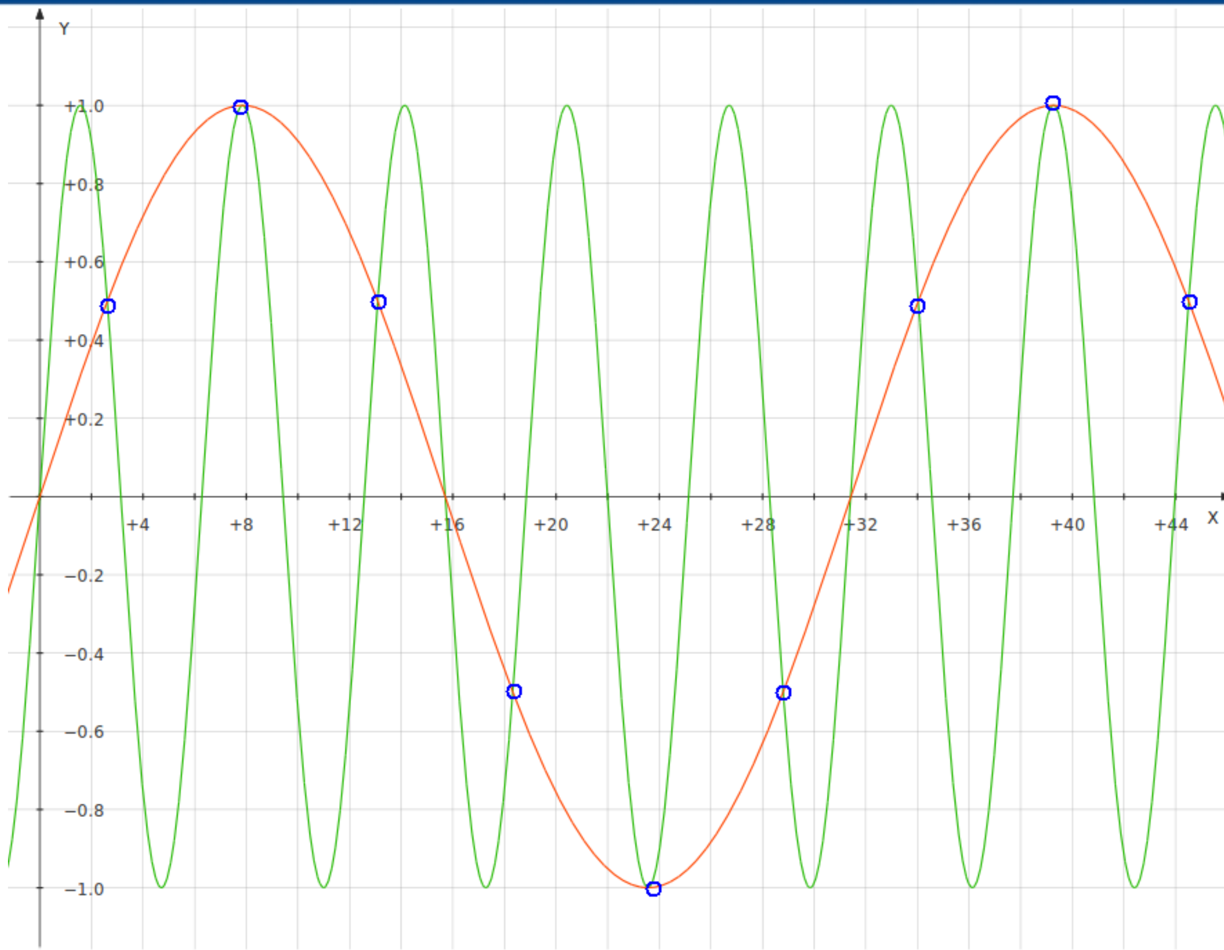


Image from: Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer. 2006

Importance of train set construction

go.mail.ru



Usually:

- random sampling is not the best strategy
- labeling is very expensive

Modern books about ML:

- Pattern Recognition and Machine Learning — 0
- Elements of Statistical Learning — 0
- An Introduction to Information Retrieval — 1 paragraph

Book about TS construction:

Burr Settles. *Active learning.*

- Publication Date: 2 July 2012

Idea of Active Learning

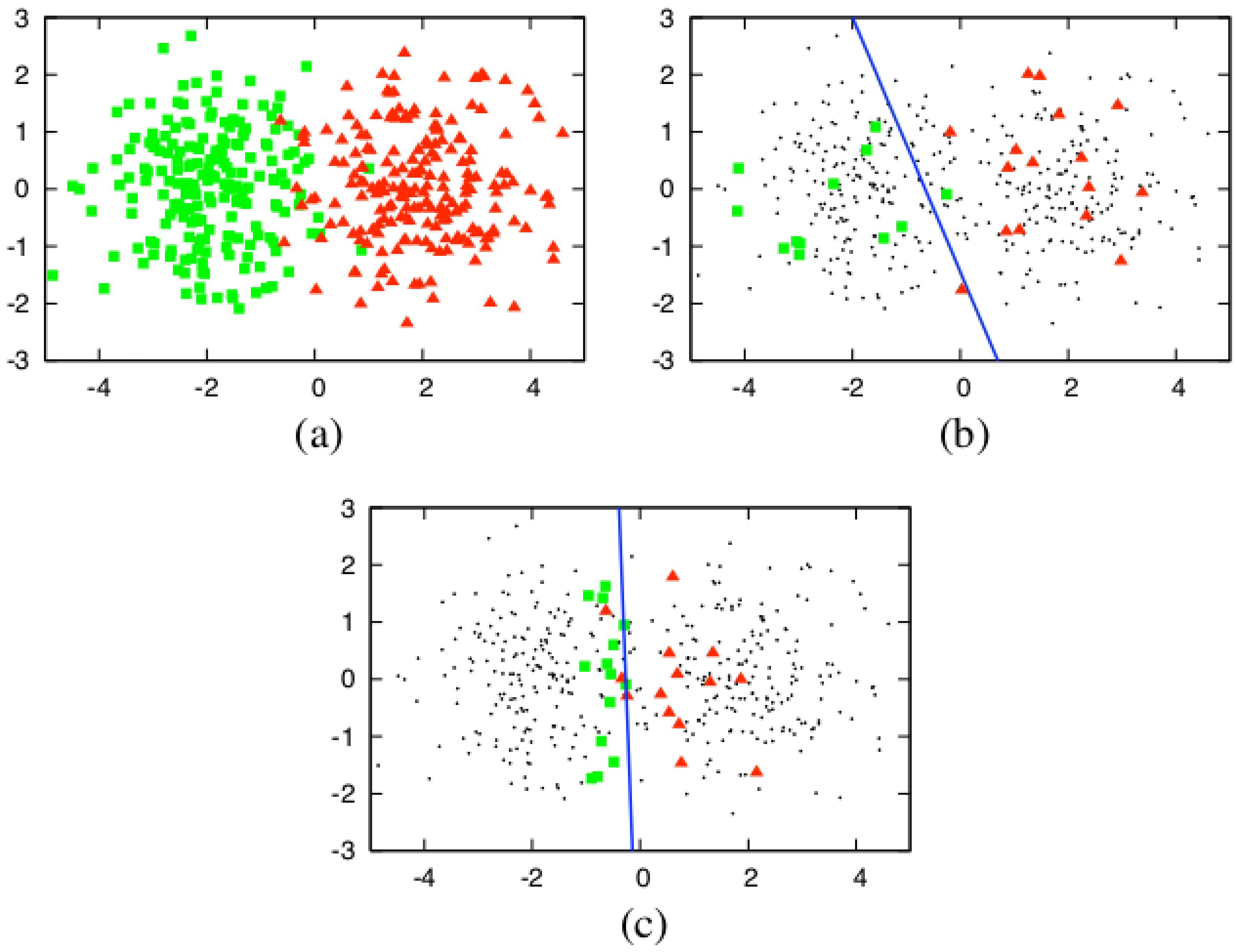


Image from: Burr Settles. *Active Learning Literature Survey.* Computer Sciences Technical Report 1648, University of Wisconsin–Madison. 2009.

Active learning workflow

Main assumption: obtaining an unlabeled instance is free

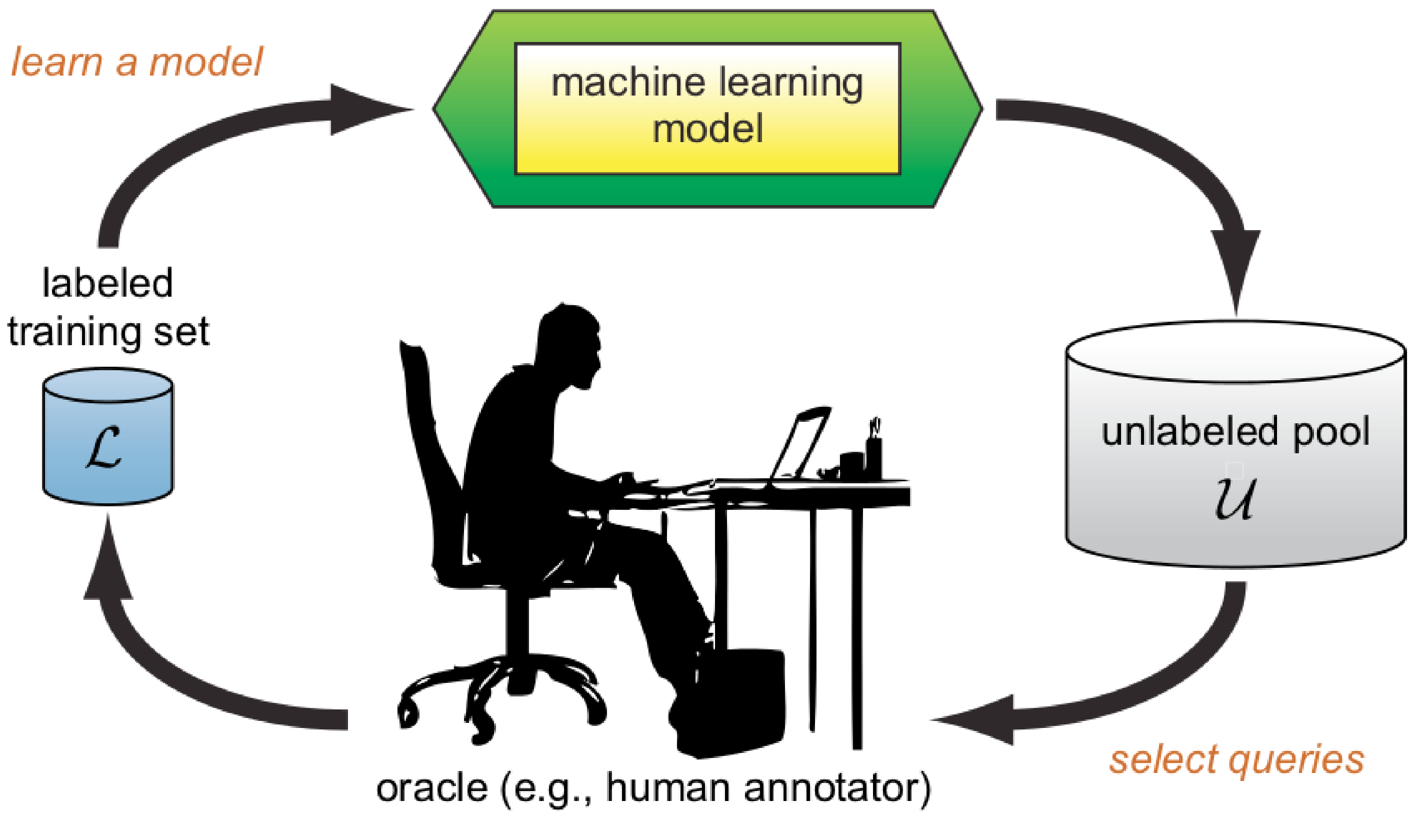


Image from: *Burr Settles. Active Learning Literature Survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison. 2009.*

Sampling scenarios

go.mail.ru

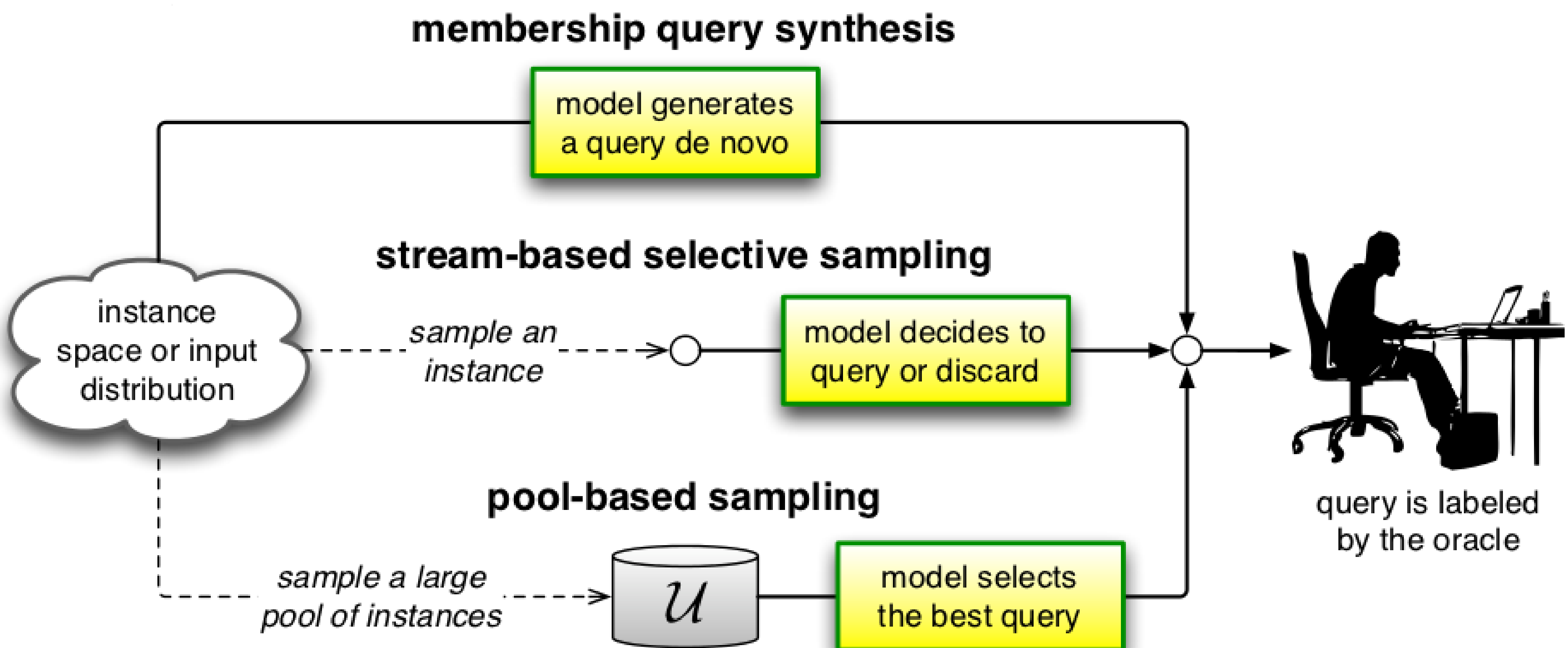


Image from: *Burr Settles. Active Learning Literature Survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison. 2009.*

Uncertainty Sampling

Take instances about which it is least certain how to label.

Least confident:

$$x^*_{LC} = \underset{x}{\operatorname{argmax}}\ 1 - P_\theta(\hat{y}|x)$$

Margin sampling

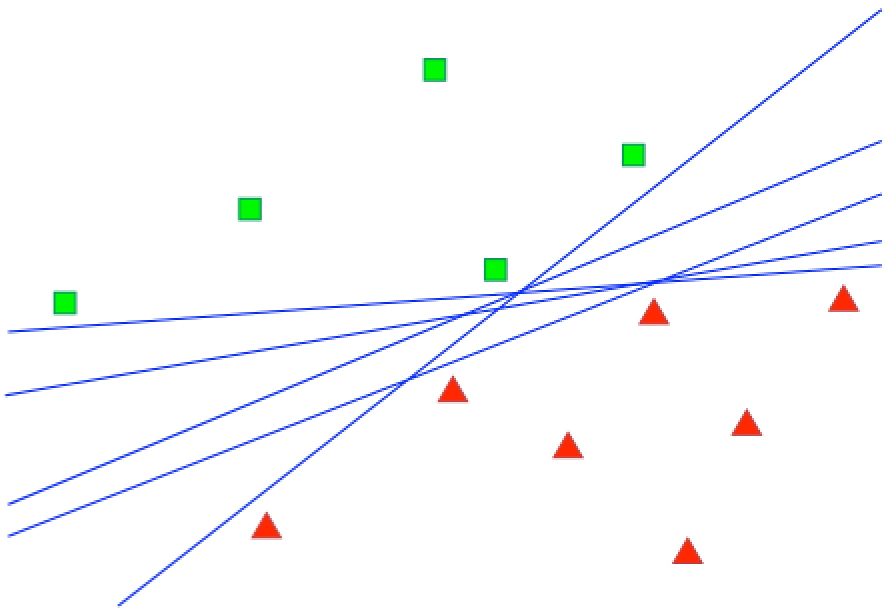
$$x^*_{M} = \underset{x}{\operatorname{argmin}}\, P_\theta(\hat{y}_1|x) - P_\theta(\hat{y}_2|x)$$

Entropy

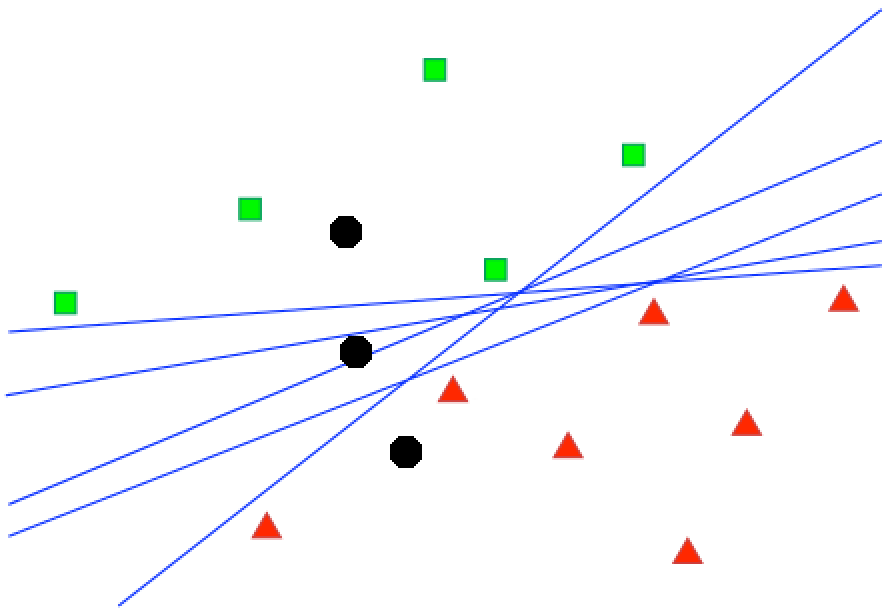
$$x^*_{H} = \underset{x}{\operatorname{argmax}} - \sum_i P_\theta(y_i|x) \log P_\theta(y_i|x)$$

Query-By-Committee

go.mail.ru

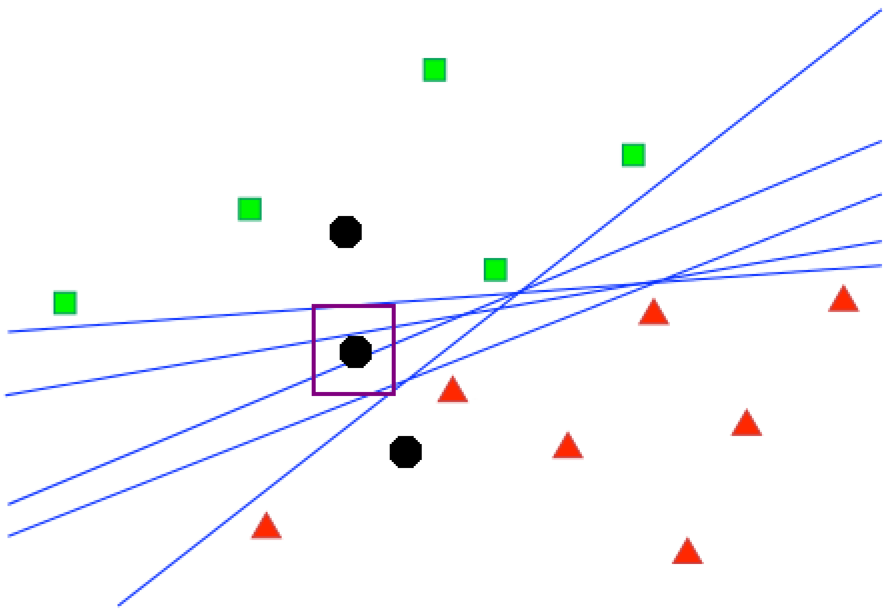


Query-By-Committee



Query-By-Committee

go.mail.ru



Query-By-Committee

Measure of committee disagreement:

Vote entropy

$$x^*_{VE} = \underset{x}{\operatorname{argmax}} - \sum_i \frac{V(y_i)}{C} \log \frac{V(y_i)}{C}$$

Kullback-Leibler (KL) divergence:

$$x^*_{KL} = \underset{x}{\operatorname{argmax}} \frac{1}{C} \sum_{c=1}^{C} D(P_{\theta^{(c)}} \| P_{\mathcal{C}}),$$

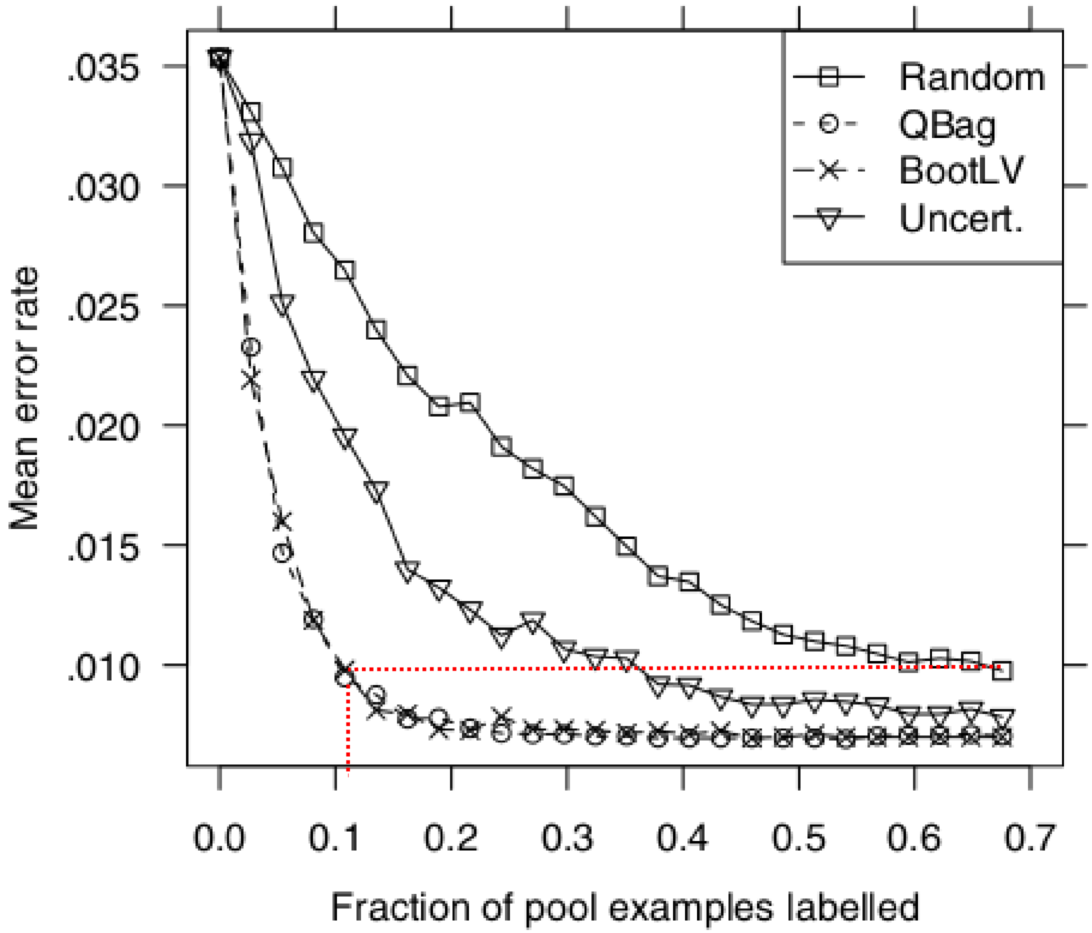
$$D(P_{\theta^{(c)}} \| P_{\mathcal{C}}) = \sum_i P_{\theta^{(c)}}(y_i|x) \log \frac{P_{\theta^{(c)}}(y_i|x)}{P_{\mathcal{C}}(y_i|x)}$$

Query-By-Committee (QBag)

Input: T – labeled train set
C – size of the committe
A – learning algorithm
U – set of unlabeled objects

1. Uniformly resample T, obtain $T_1 \ldots T_C$, where $|T_i| < |T|$
2. For each T_i build model H_i using A
3. Select $x^* = \min_{x \in U} \left| \, |H_i(x)=1| - |H_i(\mathrm{x})=0| \, \right|$
4. Pass x^* to oracle and update T
5. Repeat from 1until convergence

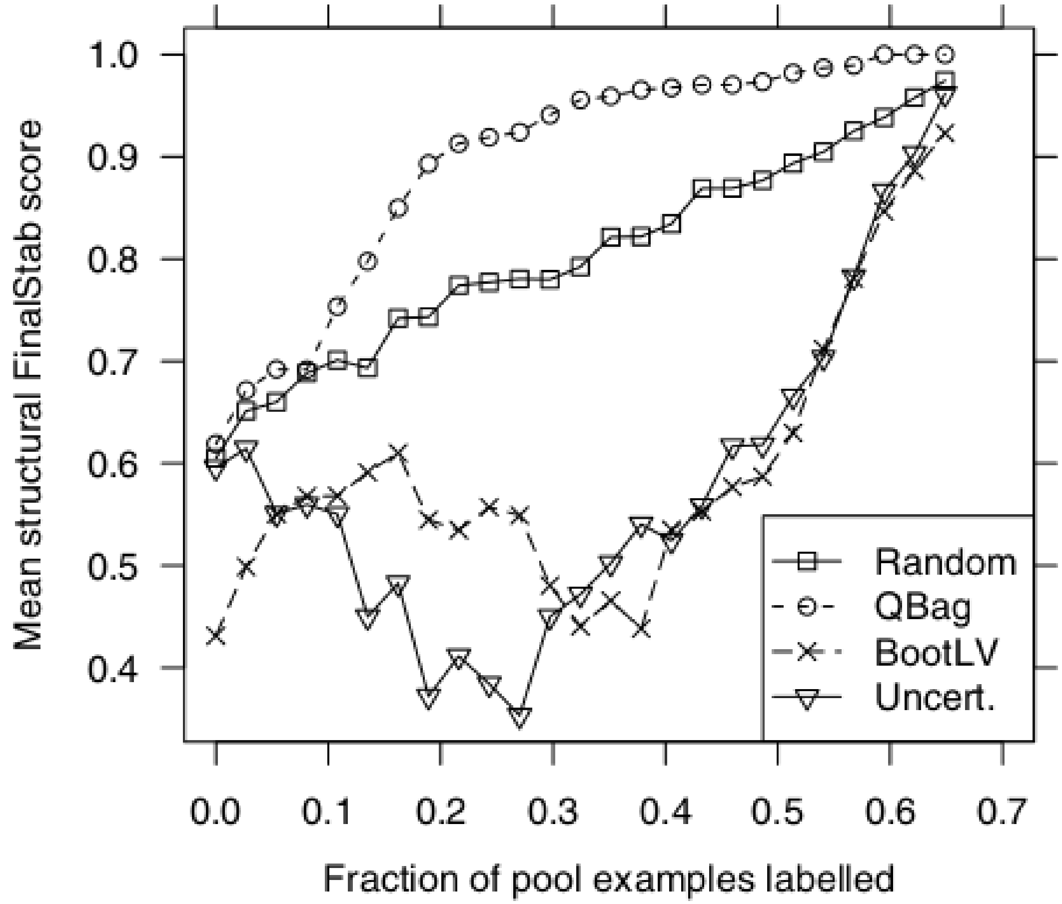
QBag: Quality



(b) Error rate

K.Dwyer, R.Holte, *Decision Tree Instability and Active Learning, 2007*

QBag: Stability



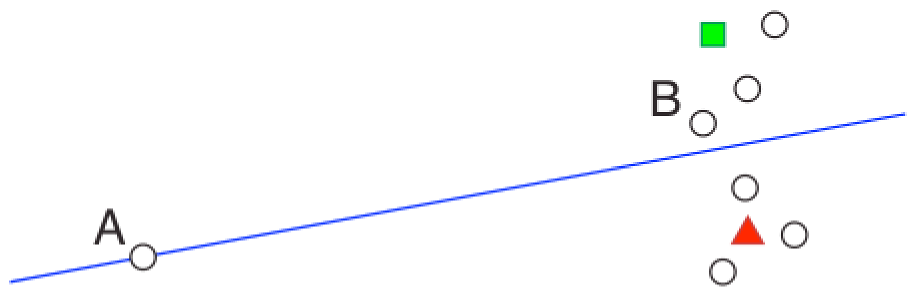
(a) Structural FinalStab ($\epsilon = 0$)

K.Dwyer, R.Holte, *Decision Tree Instability and Active Learning, 2007*

Density-Weighted Methods

go.mail.ru

Idea: Inhabit dense/sparse regions of the input space



$$x^*_{ID} = \underset{x}{\operatorname{argmax}}\, \phi_A(x) \times \left(\frac{1}{U} \sum_{u=1}^{U} \operatorname{sim}(x, x^{(u)}) \right)^{\beta}$$

Other methods

Expected Model Change

Expected Error Reduction

Variance Reduction

Many «model dependent» methods

Example 1:
Recognition of sentence boundaries

«I recommend Mitchell (1997) or Duda et al. (2001). I have strived...»

Idea:

1. Make set of simple regexp-based rules like
«*big letter at the right*» or
«*digits at the rigth and left*»
(40 rules)

2. Build classifier:
Gradient Boosted Decision Trees

A.Kudinov, A.Voropaev, A.Kalinin. A HIGHLY PRECISIONAL METHOD FOR THE RECOGNITION OF SENTENCE BOUNDARIES. Computational Linguistics and Intellectual Technologies. 2011

Example 1:
Recognition of sentence boundaries

	OpenNLP / Sentence Boundary Detector / Random sampling	«AOT» project	Mail.Ru detector / Random sampling	Mail.Ru detector / least confident
Error Rate	41,2 %	30,4 %	8.2 %	0,8 %

Train set: 9820 examples
Validation: 500 examples

Example 2: Ranking formula

go.mail.ru

Idea:

1. Query-Document presented by x, where:

   x_1 – tf-idf rank

   x_2 – query geographical

   x_3 – geo-region of user is equal to document's region

   (600 other factors)

2. Build train set:

   5 – vital

   4 – exact answer

   3 – usefull

   2 – slightly usefull

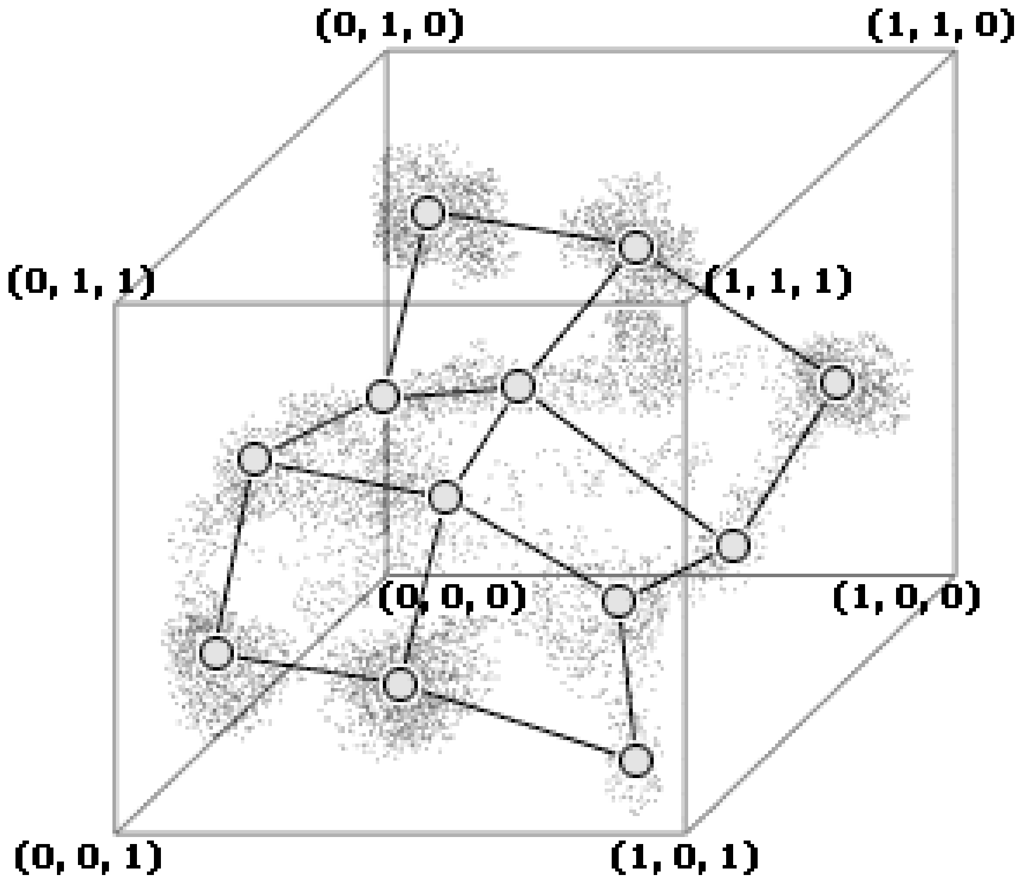
   1 – out of topic

   0 – can't be labeled (unknown language, etc.)

3. Train ranking formula using modified LambdaRank

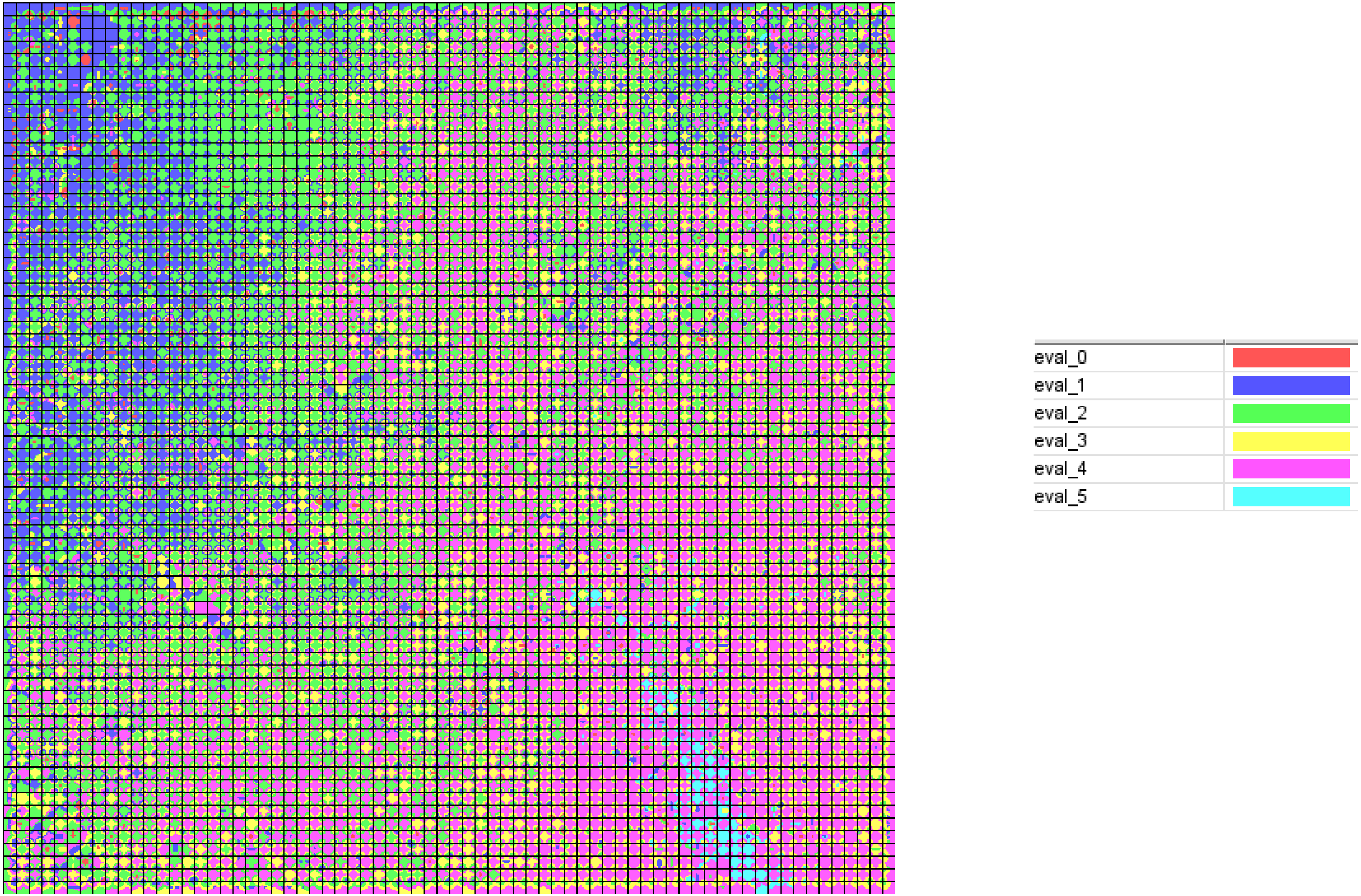
Example 2:
Self-organizing map

Idea: mapping from N-dimentional to 2-dimentional



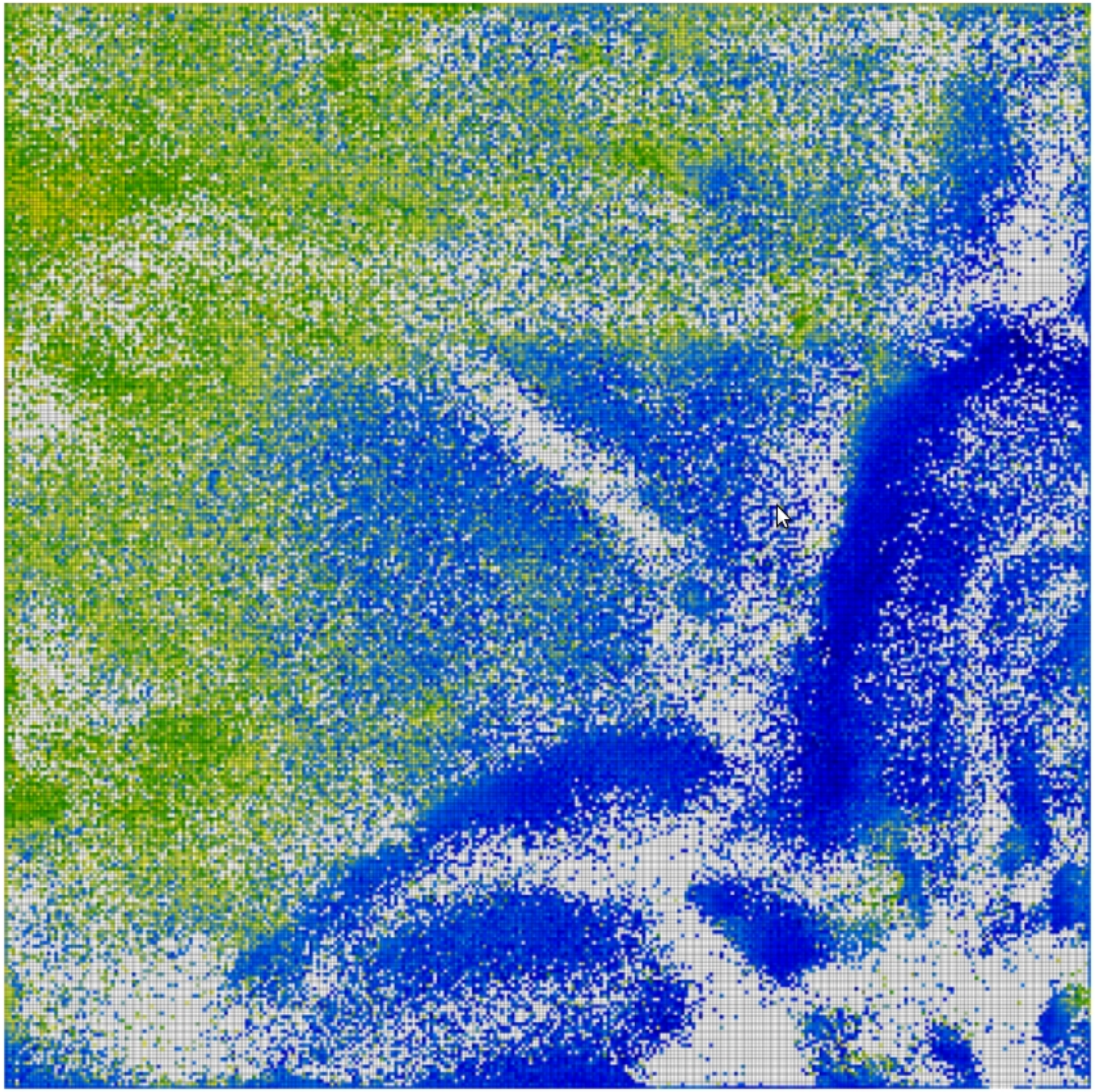
Example 2:
Map of train set for ranking

go.mail.ru



Example 2:
Density of QDocuments on the map

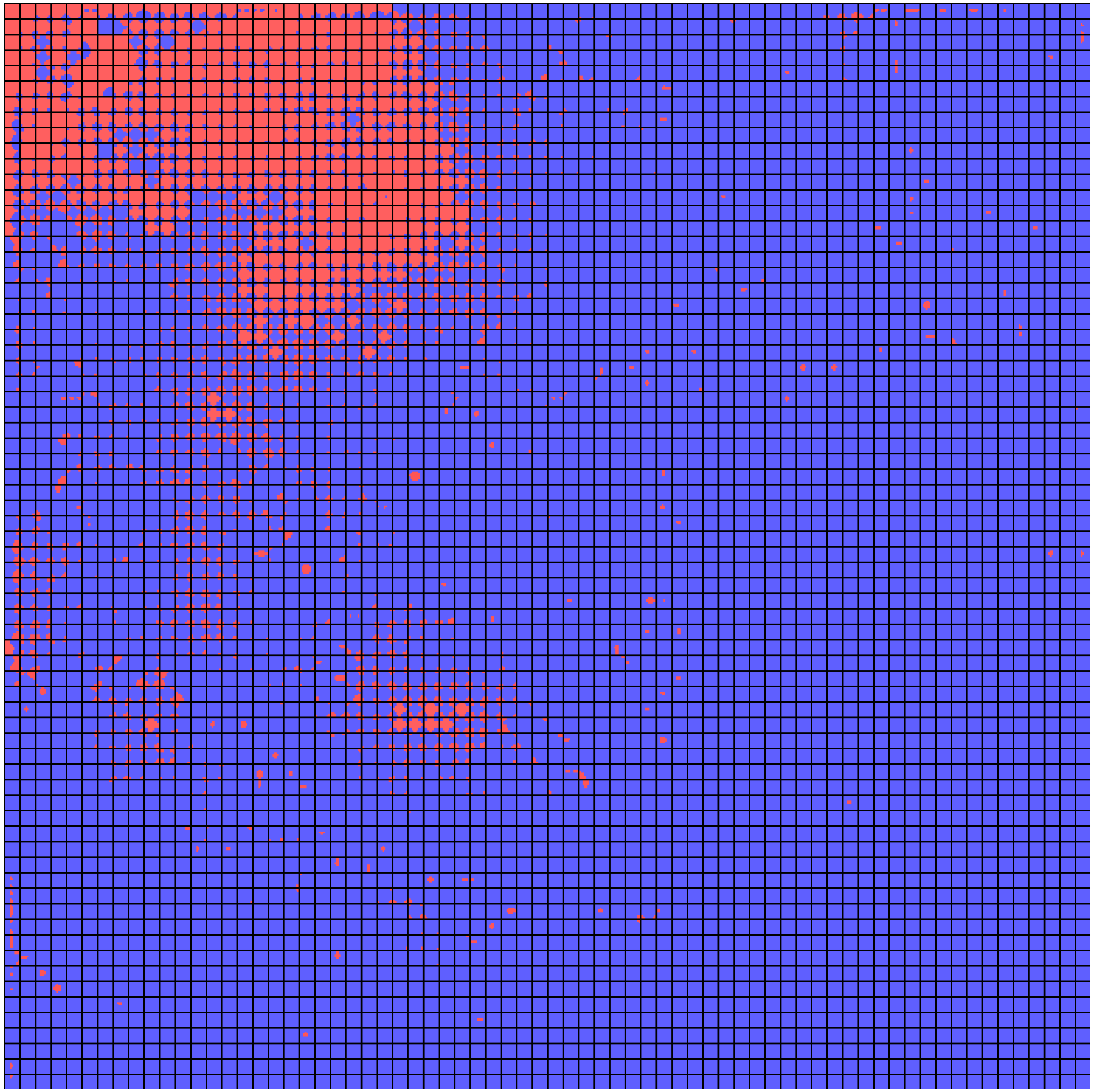
go.mail.ru



High density
Low density

Example 2:
Result of “SOM miss” selection

go.mail.ru



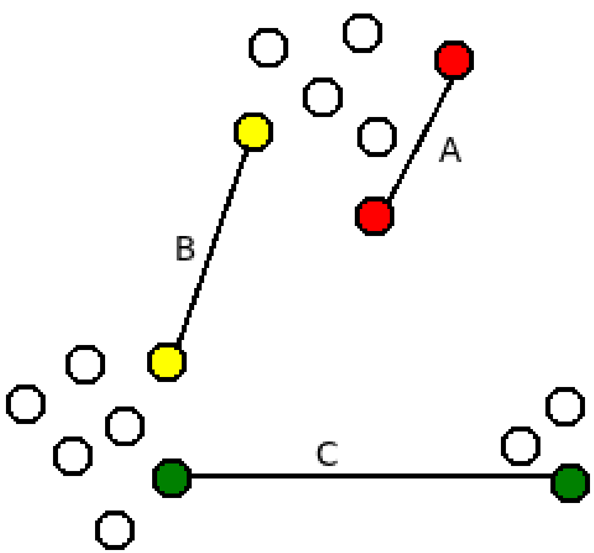
Old documents
New documents

Example 2:
Qbag + SOM heuristic

go.mail.ru

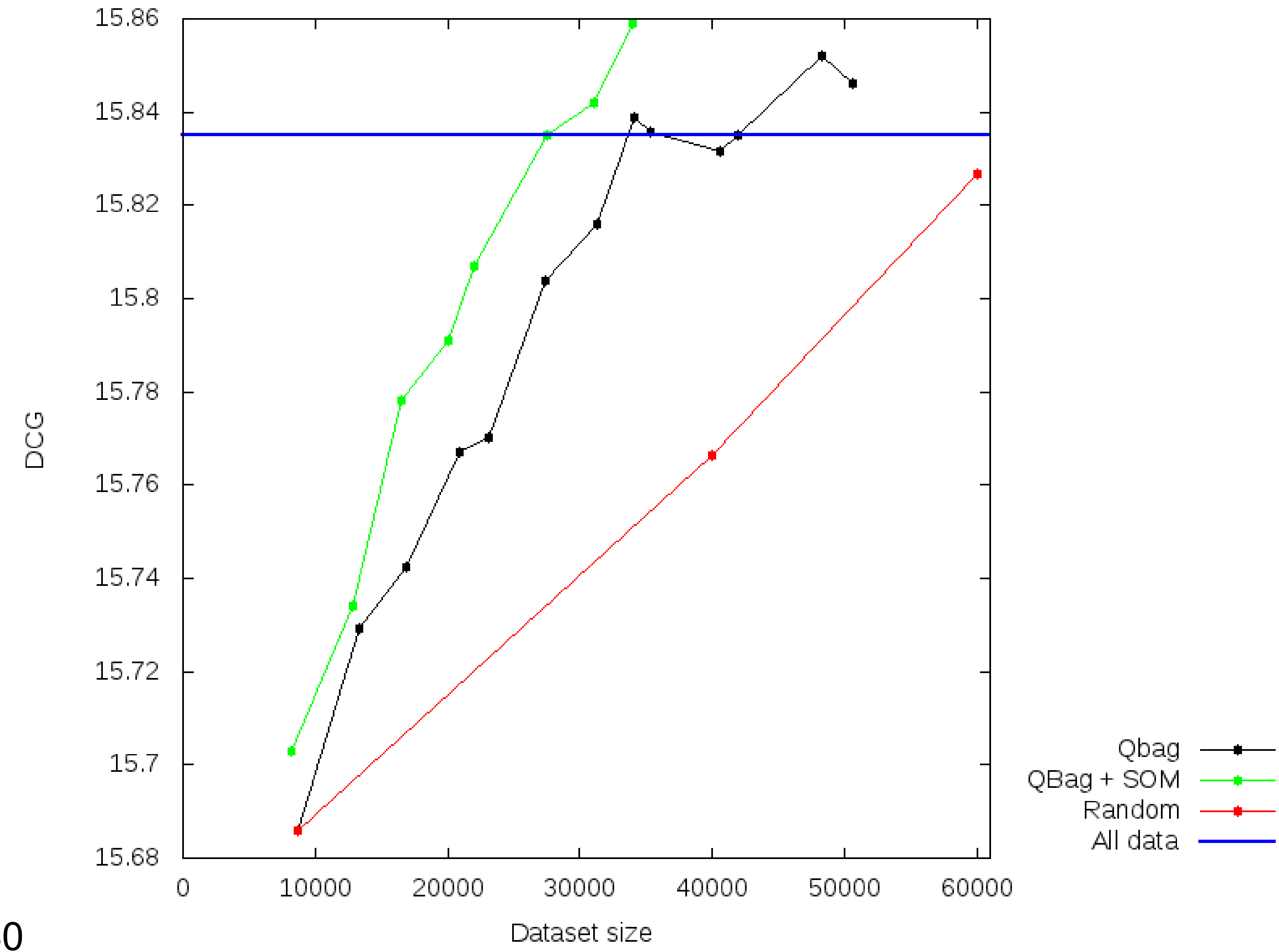
1. Transform regression to binary classification problem:
   - We should order pair of QDocuments
   - Apply *QBag* and select pairs of QDocuments

2. Apply SOM heuristic
   - Construct initial trainset *T*
   - Filter selected pairs



A) Don't take pair
B) May be
C) Take pair

Example 2:
Qbag + SOM heuristic: results

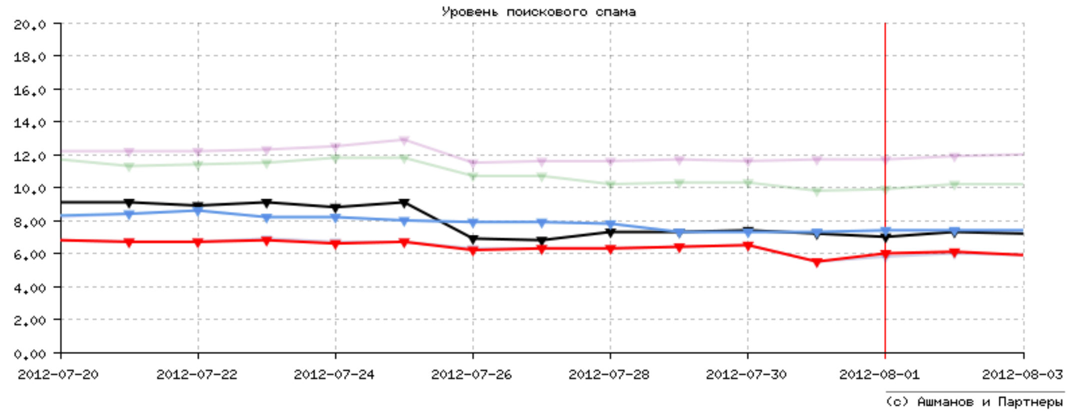


Example 3,4:
Antispam, Antiporn

go.mail.ru

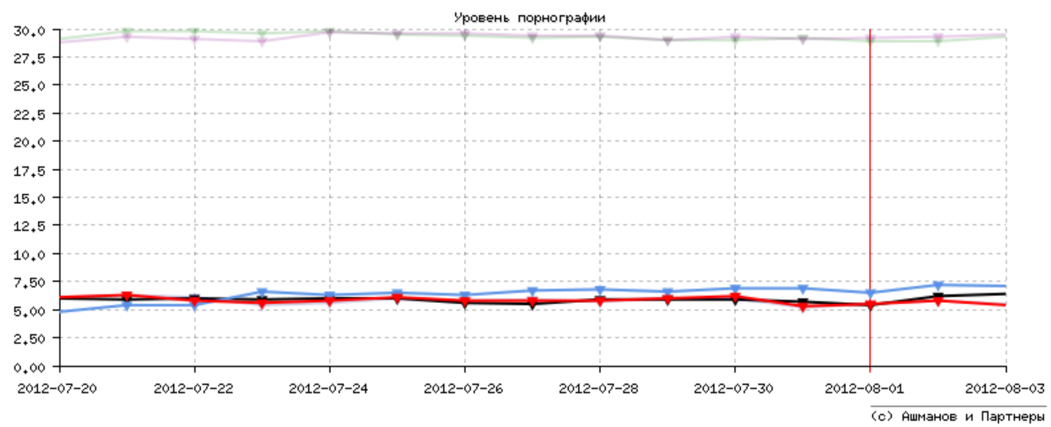
Antispam

- Neuron net
- SOM balancing



Antiporn

- Decision trees
- Uncertanty sampling



Thank you!

Reference:

- http://active-learning.net/

- Burr Settles. *Active Learning Literature Survey.* Computer Sciences Technical Report 1648, University of Wisconsin–Madison. 2009.